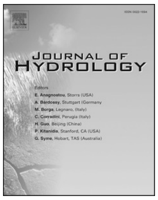
Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods

Hang Wan [a,c], Rui Xu [b,*], Meng Zhang [b], Yanpeng Cai [c,a], Jian Li [b], Xia Shen [d]

[a] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China
[b] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[c] Guangdong Provincial Key Laboratory of Water Quality Improvement and Ecological Restoration for Watersheds, Institute of Environmental and Ecological Engineering, Guangdong University of Technology, Guangzhou 510006, China
[d] Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A&F University, Yangling, Shaanxi 712100, China

ARTICLE INFO

This manuscript was handled by Huaming Guo, Editor-in-Chief

Keywords:
Non-point source pollution
Spatial characteristics
LSTM
VGG

ABSTRACT

Non-point source (NPS) pollution is an important factor affecting the quality of water environment. In recent years, a large number of online water quality monitoring stations have been used to obtain continuous time series water quality monitoring data. These data provide the necessary basis for the application of deep learning methods in water quality prediction. However, the prediction accuracy of traditional deep learning methods is low, especially in predicting the water quality with NPS pollution. Aiming to address this limitation, a novel deep learning model named SOD-VGG-LSTM with the simulation-observation difference (SOD) modular based on physical process, the visual geometry (VGG) modular reflecting spatial characteristics, and the long short-term memory (LSTM) modular based on deep learning method was developed to improve the accuracy of the water quality prediction with NPS pollution. The established model can overcome the problem that mechanism models can not predict the changes of water quality on the hourly or minute time scale. The model was applied in Lijiang River watershed. Experimental results indicated that the proposed model had the highest accuracy in the extreme value prediction compared with the mechanism model and LSTM model. The maximum relative errors between the predicted and observed results for DO, COD_{Mn}, NH_3-N, and TP were 8.47%, 19.76%, 24.1%, and 35.4%, respectively. The model evaluation demonstrated that the established SOD-VGG-LSTM model achieved superior computational performance compared to Auto Regression Integreate Moving Average model (ARIMA), Support Vector Regression model (SVR), and Recurrent Neural Network model (RNN). The evaluation results showed that SOD-VGG-LSTM achieved 3.2–39.3% higher R^2 than ARIMA, SVR and RNN. The proposed model can provide a new method for water quality prediction with NPS pollution.

1. Introduction

Non-point source (NPS) pollution is one of the important factors causing water quality deterioration (Dong et al., 2018; Xie et al., 2018). Accurate prediction of water quality changes caused by NPS pollution is of great significance to regional water environment protection. However, NPS pollution is characterized by randomness and uncertainty with complex transport process and mechanism (Huang et al., 2016; Zuo et al., 2021), which lead to inconvenience in simulating and predicting water quality changes caused by NPS pollution. Therefore, it is imperative to develop an effective and accurate water quality prediction model to simulate and predict the change of water quality caused by NPS pollution.

Mechanism models based on physical process have been developed and used to predict water quality changes caused by NPS pollution for many years (Paparrizos and Maris, 2017), such as SWAT (Soil & Water Assessment Tool), HSPF (Hydrological Simulation Program - FORTRAN) and MIKE SHE (MIKE System Hydrological European) models. Among them, SWAT model might be the most popular and concerned model for water quality prediction with NPS pollution at present, but problems of complex structures, redundant parameters and uncertain assumptions are difficult to ignore (Bahman et al., 2018). HSPF model required high-precision data. For areas with low-precision data, the error of prediction results was large (Liu and Tong, 2015). As a commercial software,

[*] Corresponding author.
E-mail address: xur@guet.edu.cn (R. Xu).

https://doi.org/10.1016/j.jhydrol.2022.128081
Received 16 February 2022; Received in revised form 5 June 2022; Accepted 6 June 2022
Available online 18 June 2022

MIKESHE did not support secondary development. The assumptions in the model might have inestimable differences for different regions and were difficult to correct (Wan et al., 2021a). Mechanism models for water quality prediction are usually constructed based on understanding the physical processes and factors. The advantage of mechanism models is that parameters of those models have strict physical interpretation (Zhou et al., 2021), but problems of difficulties in parameter calibration, complex modeling structures, uncertain model parameters and high calculation cost also limit their application in watershed water quality prediction. Additionally, mechanism models are difficult to calibrate, and often require high levels of expertise to implement. Moreover, typical applications of mechanism models for close to real time or short time in water quality caused by NPS pollution are limited (Senent-Aparicio et al., 2019).

Compared with mechanism models, deep learning models may be effective tools to overcome those limitations (Cui et al., 2016; Tiyasha et al., 2020; Zhang et al., 2021a). Deep learning models can handle nonlinear and highly stochastic predictions through dynamically and adaptively correcting model elements, which can effectively reduce the workload of modelers. And, different from mechanism models, deep learning models just focus on the input–output relationship without considering the causal relationship between parameters, which can bring convenience to the modeling process for environmental managers. Therefore, deep learning models may replace mechanism models to some extent and become effective tools for water quality prediction in the future. Methods based on the back propagation (BP) network and radial basis function (RBF) neural network could provide certain applicability in water quality prediction (Wang et al., 2013), but the problem of insufficient training could not be ignored (Deng et al., 2021). The hybrid model combined with mechanism model and artificial neural network (ANN) was developed to improve the accuracy of water quality prediction, but the model could not learn the state characteristics between time series data, which could result in large errors in extreme value prediction (Navideh et al., 2020). Recently, a new type neural network namely long short-term memory (LSTM) neural network that considered the long-term dependence in time series data has been introduced into the field of water quality prediction to improve the accuracy of extreme value prediction (Nitzan et al., 2021; Jiang et al., 2021). Previous studies have shown that compared with traditional neural network models, LSTM model is more accurate and suitable for time series data prediction (Nitzan et al., 2021; Wan et al., 2021b; Xu et al., 2021). However, LSTM model cannot reflect the impact of spatial characteristics on NPS pollution in study area. Fortunately, the convolutional neural network (CNN) was developed and applied to extract spatial characteristics due to its strong image recognition performance (Krizhevsky et al., 2017; Shelhamer et al., 2017), which can reflect the impact of spatial characteristics on NPS prediction (Chen et al., 2016; Baek et al., 2020). As a representative of CNN model, the visual geometry group (VGG) model has stable performance and concise structure (Mcilwaine and Rivas, 2020), which can be an effective method to reflect the impact of spatial characteristics on NPS pollution with deep learning methods. These studies show the potential application of the hybrid model that couples mechanism model, VGG model and LSTM model in water quality prediction caused by NPS pollution.

In the paper, a hybrid deep learning model with the simulation-observation difference (SOD) modular, the VGG modular and the LSTM modular was developed to predict the water quality changes caused by NPS pollution in watershed. The developed model has the following innovations (a) a intelligent model coupling with deep learning and feature extraction methods was developed to reflect the impact of spatial features on water quality in the study area, (b) a method to estimate the hourly water quality, obtained by combining mechanism model and intelligent model, was proposed and then verified by comparison with observation results in Lijiang River watershed, (c) the established method can improve the prediction accuracy of extreme value for water quality with coupling mechanism method and deep learning methods, (d) the developed model was applied to Lijiang River watershed and performed very well in water quality prediction.

2. Problem definition

The change of water quality in watershed is the result of comprehensive actions of multiple factors, such as spatial factors and meteorological factors (Wijesiri et al., 2015; Hu et al., 2020). However, traditional LSTM models cannot reflect the impacts of spatial factors on water quality in watershed. Because LSTM models do not have the ability to identify and extract spatial features of the watershed. The research on coupling spatial features into deep learning models to predict water quality changes caused by NPS pollution has not been carried out, and faces some challenges (Huang and Simon, 2002; Bahaa et al., 2012; Wan et al., 2021b). Fortunately, VGG methods can extract spatial features and reduce data dimension through convolution processes, which brings opportunities to predict water quality changes caused by NPS pollution by coupling spatial features into deep learning method. Hence, a novel intelligent model based on LSTM and VGG methods was developed to overcome those limitations for water quality prediction caused by NPS pollution.

3. Research area and data collection

The rapid development of planting industry and urbanization in the study area has affected the water quality of Lijiang River. Research have shown that the total amount of COD and NH_3-N entering Lijiang River through runoff in Yangshuo County alone has reached 12766 t/a and 2553 t/a (Xu et al., 2010). The concentrations of NO_3^- at the source of Lijiang River was between 2.16 and 3.32 mg/L, while that of Guilin raised to 14.35 mg/L (Li et al., 2019). The locations of study area and Stations' distribution along the Lijiang River are shown in Fig. 1. The hybrid deep learning model coupling SOD modular, VGG modular and LSTM modular was established to improve the accuracy of water quality prediction with NPS pollutants in the paper.

There are 16 hydrometeorology and water quality monitoring stations in the study area. Data collected from monitoring stations include the hydrometeorology data and pollutant data. Hydrometeorology data include conductivity (EC), Potential of hydrogen (PH), turbidity (TB), flow rate (Q), water temperature (WT), and rainfall (PCP); pollutant data include total phosphorus (TP), chemical oxygen demand (COD_{Mn}), ammonia nitrogen (NH_3-N), and dissolved oxygen (DO). Spatial data includes land-use, vegetation, and slope. Details are shown in Table 1. The water quality of Yangshuo Station for next time steps was predicted based on the antecedent meteorological, hydrological, water quality data at neighboring 15 stations and spatial data in the study area. The locations of these stations were shown in Fig. 1. These hourly data comprising 4 years from 2016.04 to 2019.12 were divided into training and testing sets chronological order, each accounting for 80% and 20% of the total data respectively. The training set was used to train the model parameters and testing set was employed to evaluate the model performance.

4. Methods

4.1. Framework

A hybrid deep learning model coupling with SOD modular, VGG modular and LSTM modular was developed in the paper. The model frame is shown in Fig. 2, including data collection part, VGG feature extraction part, pollutant transport and diffusion simulation part, error correction part, and result analysis part. Firstly, the research data, including spatial information, hydrometeorology parameters, and pollutant parameters, were collected and pre-processed. Secondly, the VGG model was adopted to extract the spatial features of the watershed to generate the multi-dimensional time-series data with spatial features.

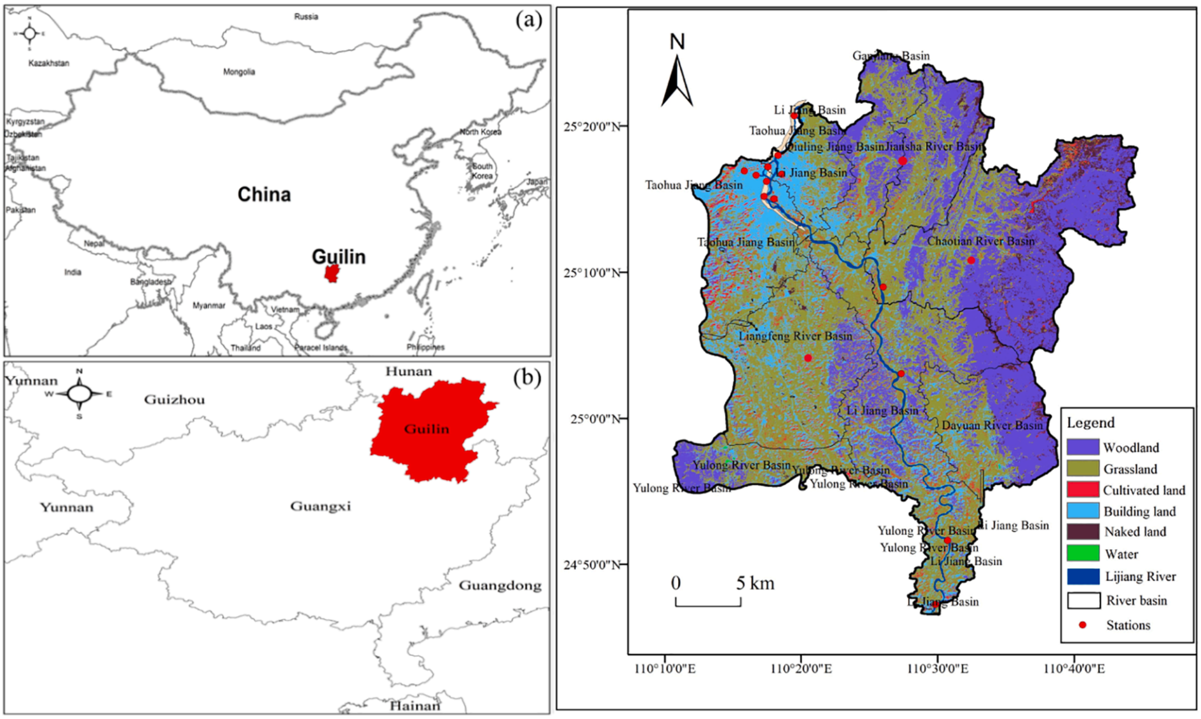


Fig. 1. Overview of the study area.

Table 1
Pollution, hydrometeorology, and spatial data parameters.

Data category	Parameter	Resolution	Unit	Source
Hydrometeorology (2016–2019)	EC	Hourly	μs/cm	Guilin Ecological environment Bureau
	PH	Hourly	Dimensionless	
	TB	Hourly	NTU	
	Q	Hourly	m^3/s	
	WT	Hourly	°C	
	PCP	Hourly	mm	
Water quality (2016–2019)	TP	Hourly	mg/L	Guilin Ecological environment Bureau
	TN	Hourly	mg/L	
	COD_{Mn}	Hourly	mg/L	
	NH_3-N	Hourly	mg/L	
	DO	Hourly	mg/L	
Vegetation (2016–2019)	NDVI	16-day	Dimensionless	Earth Explorer

Thirdly, the spatial correlation of water quality monitoring stations was studied, and multi-source water diffusion model was built. The difference between the simulated and observed results was used to construct the error sequence. Fourthly, a hybrid deep learning model was developed to predict water quality changes caused by NPS pollution. Finally, using the root-mean-square error (RMSE), mean absolute error (MAE), and symmetric mean absolute percentage error (SMAPE) as the evaluation parameters, the model parameters are optimized through experiments. The accuracy of the proposed model is verified through comparison with state-of-the-art prediction models.

4.2. Feature extraction through VGG

Water quality changes are affected by spatial factors, such as topography, vegetation, and slope, and have a nonlinear trend. The watershed considered in this study is dominated by planting and has high vegetation coverage, river channels, and distinct geographical features of the sub-watershed. The middle and northeastern regions of the watershed are flat; the north and northwest regions are the origins of the main river channels, which have alpine forest topography; the other regions have typical karst topography. In this study. VGG model was used to determine the influence of spatial features on water quality through convolution processes. The model includes 13 convolutional layers, 3 fully connected layers and 5 pooling layers. A 3*3 Convolution Kernel was adopted to extract image features by scanning the image matrix. As shown in Fig. 3, the original matrix data (image) can be scanned through a fixed matrix (Convolution Kernel) to obtain image features (Simonyan and Zisserman, 2004). Results in Fig. 3 indicated that the features of the original image could be extracted and the dimension of the image could be reduced by the convolution processes. The general law can be summarized that the matrix dimension after convolution process is equal to the difference between the matrix dimension of image and the matrix dimension of convolution kernel plus 1.

The convolution layer has a continuous 3 × 3 convolution kernel and a maximum pooling size of 2 × 2. In the convolutional layer, the input of each layer denotes a small part of the output of the previous layer, and the size of this small part is the same as the size of the convolution kernel. The convolutional layer was used to analyze the small part of each upper layer to obtain more abstract spatial features. The convolutional layer is defined by *Eq.* (1). The ReLU function is used as an activation function of this layer because it can effectively avoid the problem of gradient disappearance and accelerate the training process. The ReLU function is defined by *Eq.* (2).

$$x_j^l = f\left(\sum_{i=M_j} x_j^{(l-1)} \bullet \mathrm{k}_{\mathrm{i}}\mathrm{j}^{\mathrm{l}} + \mathrm{b}_{\mathrm{j}}^{\mathrm{l}}\right) \tag{1}$$

$$f(x) = max(0, x) \tag{2}$$

In *Eqs.* (1) and (2), *l* represent the l^{th} network layer; M_j represents the receptive field output by the previous layer; *b* denotes the bias; *k* is the convolutional layer; and $f(\bullet)$ is the nonlinear activation function.

In VGG model, the pooling layers can reduce the size of the input feature to simplify the computational complexity of the neural network model. The max-pooling layer outputs the maximum value of the input features, and it can be defined as.

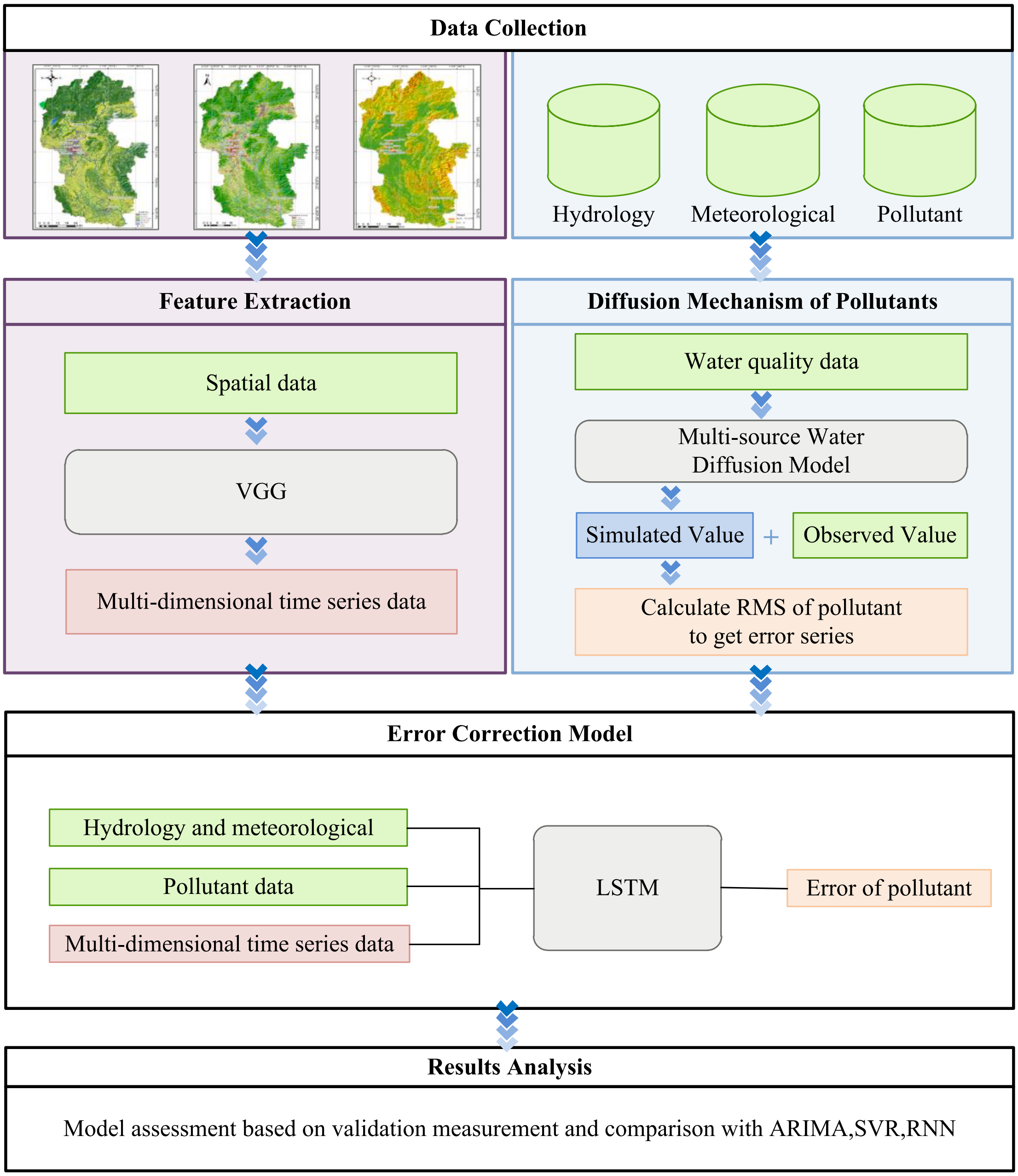

Data Collection
Hydrology
Meteorological
Pollutant
Feature Extraction
Spatial data
VGG
Multi-dimensional time series data
Diffusion Mechanism of Pollutants
Water quality data
Multi-source Water Diffusion Model
Simulated Value
+
Observed Value
Calculate RMS of pollutant to get error series
Error Correction Model
Hydrology and meteorological
Pollutant data
Multi-dimensional time series data
LSTM
Error of pollutant
Results Analysis
Model assessment based on validation measurement and comparison with ARIMA,SVR,RNN

Fig. 2. Framework of the proposed model.

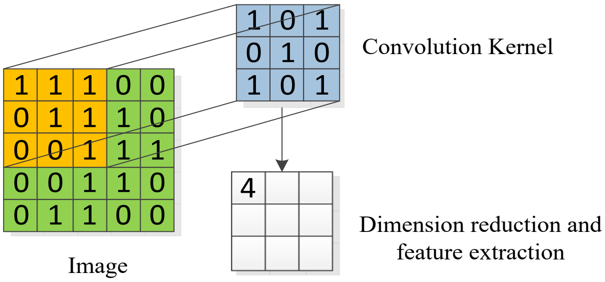


Fig. 3. Feature extraction process.

$$x_j^l = f\left(\beta_j^l p\left(x_j^{l-1}\right) + b_j^l\right) \tag{3}$$

where $p(\bullet)$ is the pooling function, and β is the weight.

The spatial remote sensing image can provide rich land attribute information, as shown in Fig. 4. Firstly, the remote sensing data, including land-use, vegetation, and slope, was collected from web. Secondly, images of spatial features were converted into an image format of224 × 224 × 3, and then, features were extracted by VGG model. Finally, the time-series high-dimensional features of spatial images were obtained. In order to eliminate the influence of noise caused by high-dimensional redundant features, the principal component analysis was used to reduce the dimensions of high-dimensional features to obtain multi-dimensional spatial feature time-series data. Combining these data with hydrometeorology data, pollutant data, and target pollutant error, the input and output dataset of the hybrid deep learning model can be constructed.

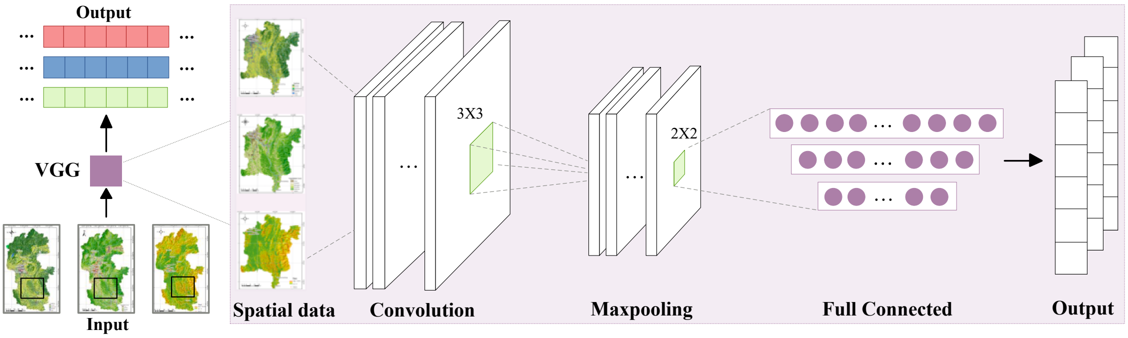


Fig. 4. Architecture of the constructed VGG model.

4.3. Simulation-observation difference for pollutant

In the paper, the mechanism model based on physical process was adopted to simulate the transport and diffusion process of pollutants in the watershed. The physical-based model is a modular structure model, which consists of hydrological process sub-module, soil erosion sub-module and pollution load sub-module. The model is mainly driven by the water balance equation, which can be expressed as.

$$SW_t = SW_0 + \sum_{i=1}^{t}(R_d - Q_{surf} - E_a - W_{seep} - Q_{gw}) \tag{4}$$

where, SW_t represents the soil water content; SW_0 represents the initial soil water content; T is the time; R_d, Q_{surf}, E_a, W_{seep} and Q_{gw} represent precipitation, surface runoff, evaporation, permeation, and underground water content, respectively.

The study area was divided into 8 sub-basins based on DEM and measured river network. The calibration and validation of the model followed the principle of first runoff followed by nutrients. The runoff was calibrated and validated by the flow data of Yangshuo Hydrological Monitoring Station, and the trend validation was carried out by the daily measured water quality data of the river in Yangshuo. Some key parameters and processes of the mechanism model were not calibrated and corrected, which could only ensure the accuracy of the model prediction trends. The purpose of model calibration only needs to ensure the accuracy of model prediction trend. Based on the predicted results, the simulation-observation difference was used to stabilize the water quality time series to reduce the prediction error of extreme value.

$$X^{Diff} = S_t - V_t,\ t = 1, \dots, N \tag{5}$$

where, N is the length of data series; S_t and V_t are the data record at time t for the simulation results and the observation results, respectively; X^{Diff} is the first-order difference based on S_t and V_t and will be used as the output for the established model. The error time series of the pollutant at the target station is obtained through X^{Diff}.

4.4. Error correlation through LSTM

Long Short-Term Memory model was used to predict the pollutant error at the next moment based on hydrometeorology indexes, pollutants indexes, spatial features indexes, and target pollutant errors in the past few moments. The rolling prediction scheme was used in this scheme. All indexes in the prediction window were used to predict pollutant error in the next moment. This process was repeated until the error prediction of the target pollutant is complete. In the simulation, the target pollutant error sequence was assumed to be E. The hydrometeorology parameters were $H\{H^1, H^2, \cdots, H^n\}$. Other pollutant indexes were $P\{P^1, P^2, \cdots, P^n\}$, and the spatial feature indexed were set as $S\{S^1, S^2, \cdots, S^n\}$. In the proposed model, the input matrix was $X\{x_1, x_2, \cdots, x_n\}$, as shown in *Eq.* (6). The target sequence was $E'\{e_1', e_2', \cdots, e_n'\}$, as shown in *Eq.* (7), where $f(X)$ denoted the training function of the LSTM neural network. The training process of LSTM model was defined by *Eqs.* (6) and (7). In the model, the size of the rolling window was *l*. The input and output of the model were respectively defined by *Eqs.* (8) and (9).

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} E_1 & H_1 & P_1S_1 \\ E_2 & H_2 & P_2S_2 \\ \vdots & \vdots & \vdots \\ E_n & H_n & P_nS_n \end{bmatrix} \tag{6}$$

$$E' = \begin{bmatrix} e_1' \\ e_2' \\ \vdots \\ e_n' \end{bmatrix} = f(X) = f\left(\begin{bmatrix} x_1 & x_2 & \cdots & x_l \\ x_2 & x_3 & & x_{l+1} \\ \vdots & & \ddots & \vdots \\ x_{n-l+1} & x_{n-l+2} & \cdots & x_n \end{bmatrix}\right) \tag{7}$$

$$X_t = [x_t, x_{t+1}, \cdots, x_{t+l-1}]^T, t \in 1, 2, \cdots, n-l+1, \tag{8}$$

$$e_t' = f([x_t, x_{t+1}, \cdots, x_{t+l-1}]), t \in 1, 2, \cdots, n-l+1. \tag{9}$$

The error series of water quality between simulated and monitored values based on mechanism model, the time series of spatial features extracted by VGG model, and hydrological and water quality data of upstream monitoring stations were used as the output or input of LSTM model to predict the pollutant concentration error of the river. Finally, the actual water quality of the river was inversed by error series of pollutant concentration predict by LSTM model and estimated values simulated by mechanism model. The final model structure was shown in Fig. 5.

4.5. Model performance evaluation

The accuracy and performance of the established SOD-VGG-LSTM model have been evaluated by a series of numerical tests in the paper. Firstly, it was examined whether the proposed SOD-VGG-LSTM model could overcome the extreme value problem of deviating data effectively. Then, the proposed model was compared with LSTM model without considering the spatial watershed features. Finally, the proposed method was compared with several state-of-the-art prediction models, namely the autoregressive moving average model (ARIMA) based on statistical theory, the support vector regression model (SVR) based on traditional machine learning, and the traditional Neural Network model (RNN) model based on deep learning.

The evaluation indexes included RMSE, MAE, and SMAPE, which indicated the deviation of the model prediction results from the true values. The evaluation indexes RMSE, MAE, and SMAPE were expressed as follows.

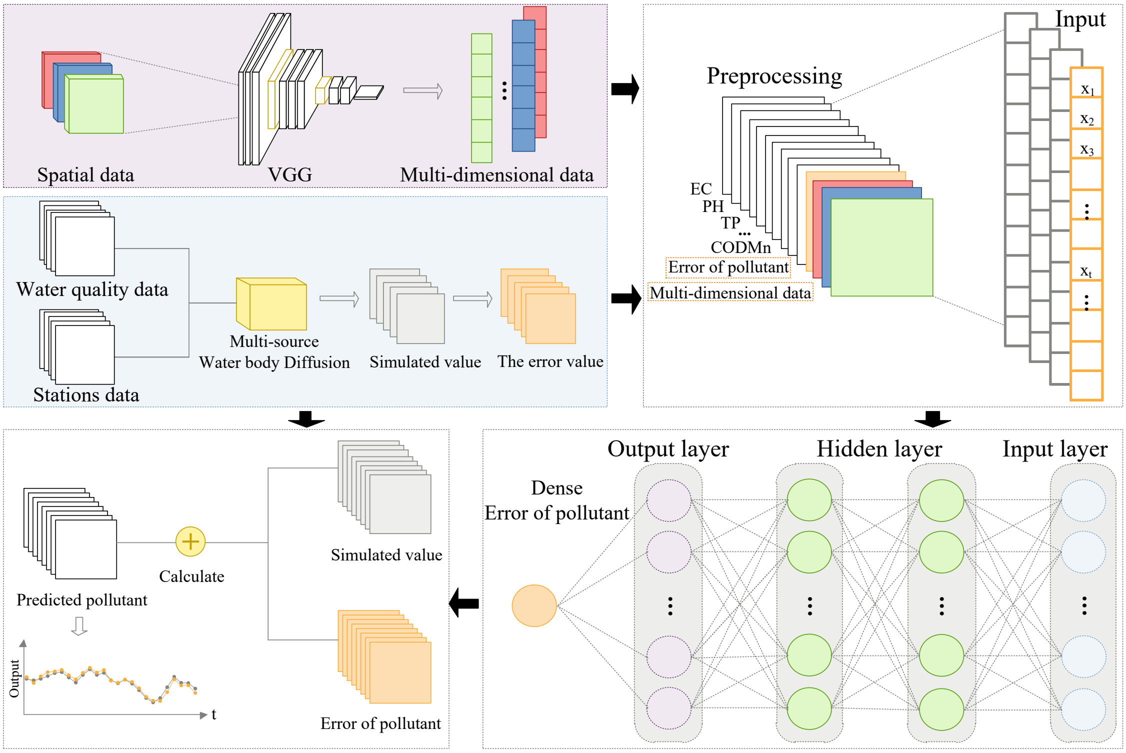


Fig. 5. Architecture of the proposed SOD-VGG-LSTM model.

$$RMSE_{(y',y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2} \tag{10}$$

$$MAE_{(y',y)} = \frac{1}{n}\sum_{i=1}^{n}|y_i' - y_i| \tag{11}$$

$$SMAPE_{(y',y)} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i' - y_i|}{(y_i' + y_i)/2}, \tag{12}$$

where n denotes the total number of samples, y_i is the observed value, and y_i' is the corresponding predicted value.

5. Results and analysis

5.1. Network parameters

Super parameters control the network structure of the LSTM model (Yan et al., 2019), which determines the simulation results. In this paper, super parameters mainly included time step and neurons. Time step represents the length of the past data used in the prediction. As the prediction process progresses, these time step values slide along the time axis in the sliding window finally reaching the end of the dataset (Xu et al., 2022). Neurons in LSTM model represent the complexity and learning ability of the established model. In water quality prediction, values of super parameters affect the performance and accuracy of the established model. Therefore, the sensitivity analysis (setting of supper parameters) needs to be evaluated by changing values of super parameters.

The performance of the model was optimized by setting super parameters. As shown in Table 2, the number of neurons was selected from the set of {16, 32, 64, 128, 256}. The test results showed that the performance of the model increased firstly and then decreased with the increase of the number of neurons. The test results indicated that the model accuracy was the highest, when the number of hidden neurons was 64. The time step values, which denoted the length of the time series required for prediction, was selected from the set of {4, 8, 12, 16, 20}. The model achieved the best prediction performance for the time step of 12. Thus, the data of the past 12 h were used to predict the concentration of pollutants at the next moment in the paper.

Table 2
Optimal parameters of the SOD-VGG-LSTM model.

Parameter	Set of feasible values	Optimal value	RMSE	MAE
Neuron number	{16, 32, 64, 128, 256}	16	1.241	0.758
		32	0.972	0.514
		64	**0.573**	**0.361**
		128	0.859	0.498
		256	1.176	0.712
Time step	{4, 8,12,16,20}	4	0.833	0.587
		8	0.602	0.412
		12	**0.468**	**0.279**
		16	0.704	0.306
		20	0.935	0.458

5.2. Spatial feature extraction

Visual Geometry Group method was adopt to extract the spatial feature and reduce the dimension of the input remote sensing image. In this way, the high-dimensional data could be characterized as multi-dimensional spatial feature time-series data, which would be used as the feature expression vectors of image data. Remote sensing data from January 2018, July 2018 and January 2019 were used to demonstrate the effectiveness of the established VGG model in this part. The data before and after dimensionality reduction were shown in Fig. 6. It could be observed that before dimensionality reduction, the original data including land-use, vegetation and slope was scattered and complex, and the amount of data was large. After dimensionality reduction, the spatial features of remote sensing image could be extracted, and the extracted data were separated, which greatly reduced the complexity of

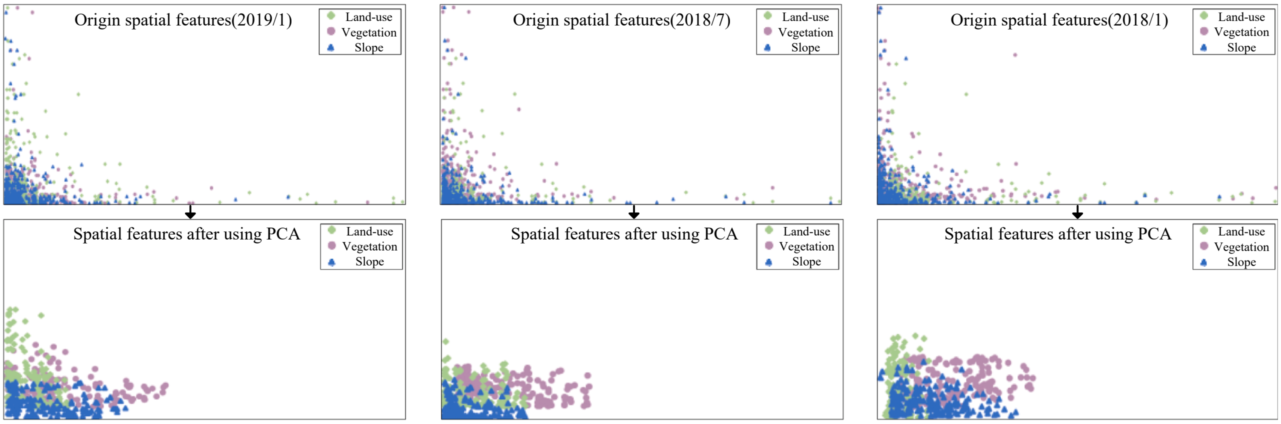


Fig. 6. Data dimensionality reduction and feature extraction.

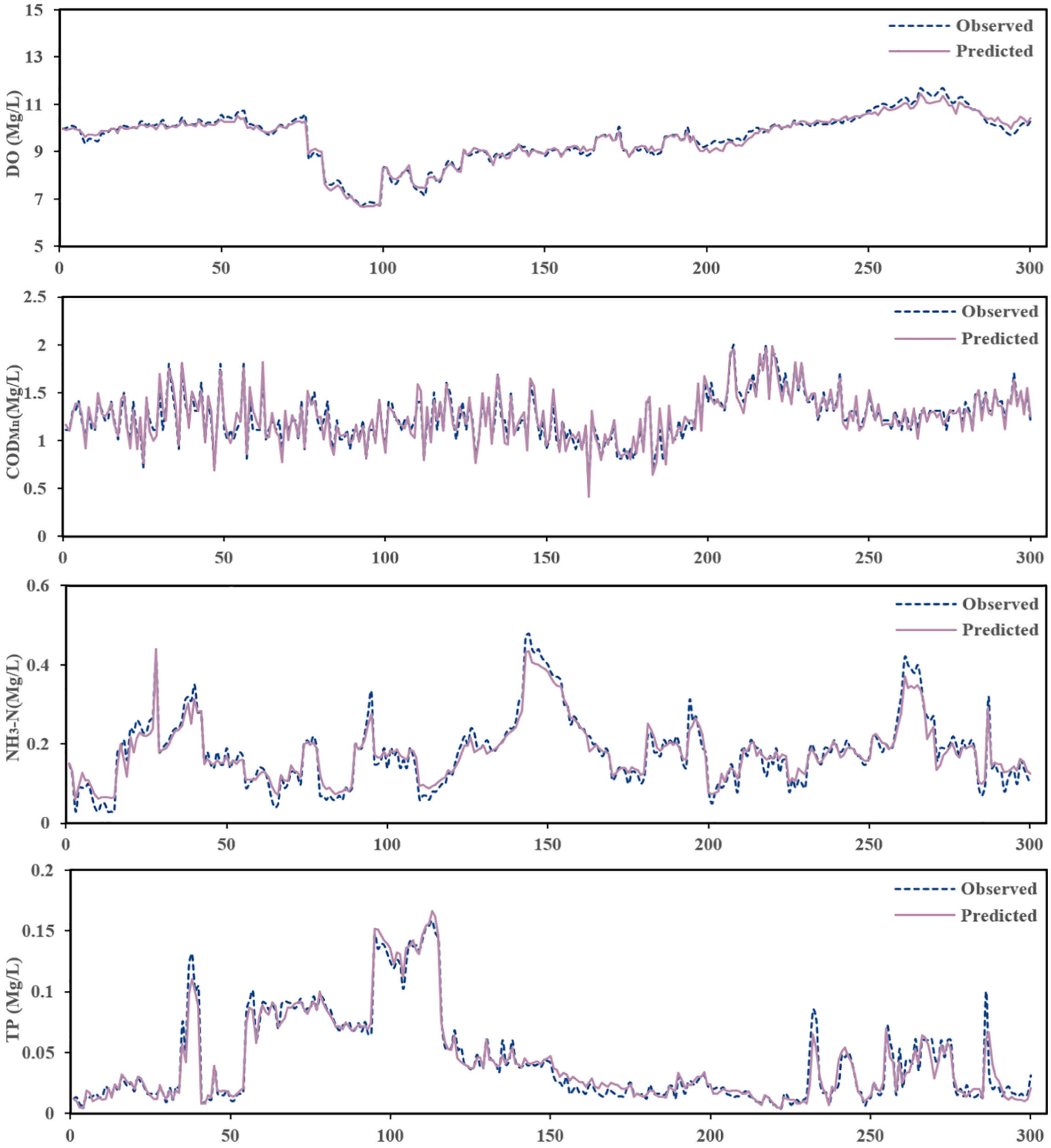


Fig. 7. The comparison of the prediction results of the SOD-VGG-LSTM model and observed values during the period from May 1, 2019 to May 12, 2019.

calculation and the identification error caused by redundant information. More importantly, through dimensionality reduction, complex remote sensing images could be transformed into simple multi-dimensional vectors, which could be directly coupled with deep learning method to realize water quality prediction with considering spatial feature.

5.3. Prediction performance for water quality

The indicators of water quality, DO, COD_{Mn}, NH_3-N, and TP, were selected as performance evaluation indicators of the established SOD-VGG-LSTM model in this part. The online monitoring data from April 2016 to March 2019 were used to train the model, and the hourly values of DO, COD_{Mn}, NH_3-N, and TP concentration from April 2019 to December 2019 were predicted. The input data were preprocessed before being put into the model. The preprocessing included deleting the wrong and missing data and replacing those with the average daily values; the 3σ principle was used to analyze and eliminate abnormal data. In addition, to improve the training speed and prediction accuracy of the model, the Z-score standardization method was used to normalize the input dataset.

After training and convergence, the optimal SOD-VGG-LSTM model was obtained. To verify the prediction performance of the optimal model, the test dataset was used for model evaluation. The comparison between the predicted values obtained by the model and the corresponding observed values during the period from May 1, 2019 to May 12, 2019 was shown in Fig. 7. The maximum relative error of DO, COD_{Mn}, NH_3-N, and TP were 8.47%, 19.76%, 24.1%, and 35.4%, respectively. The reason for the large prediction error of TP might be that some enterprises secretly emissions, leakage emissions were not considered in the established model, resulting in the sudden change of water quality.

5.4. Extreme values prediction

When NPS pollutants changes sharply under extreme meteorological conditions, traditional deep learning methods are limited by historical data and cannot effectively predict extreme values, while physical models can ensure that the predictions are within a controllable range. To evaluate the prediction performance of the proposed model for extreme values, SOD-VGG-LSTM was tested on the data that did not appear in the training dataset. The prediction results were shown in Fig. 8, which showed the comparison of the predicted, observed, and calculated results of the extreme values for DO, COD_{Mn}, NH_3-N, and TP concentration within 30 d. The prediction results showed that the maximum relative error between the predicted value and observed value were 6.07%, 11.6%, 22.1%, and 23.9%. Obviously, the established SOD-VGG-LSTM model achieved good prediction accuracy after coupling mechanism model in predicting the extreme values on a day scale.

Previous studies have indicated that spatial characteristics influenced the transport and diffusion process of NPS pollution in watershed. But traditional deep learning methods neglected the influence of spatial characteristics on the transport-diffusion process of NPS pollution. In this paper, VGG method in the established SOD-VGG-LSTM model was used to extract spatial features of the study area to improve the prediction accuracy of the model. Fig. 9 showed the comparison of the predication results for DO, COD_{Mn}, NH_3-N, and TP with 60 h obtained from SOD-VGG-LSTM model with spatial information and LSTM model without spatial information. The results of the maximum relative error between the predicted value and observed value for two models were presented in Fig. 9. The results showed that the predication accuracy of LSTM model without considering spatial characteristics is lower than that of SOD-VGG-LSTM model.

5.5. Comparison with other models

Traditional models including ARIMA, SVR, RNN were used to verify the superiority and advancement of the established SOD-VGG-LSTM model in this part. Among them, ARIMA based on statistical theory is often used to predict time series, which is a short-term prediction method with high prediction accuracy. An ARIMA model is characterized by 3 terms, *p*, *d*, *q*. where, *p* is the order of the AR term; *q* is the order of the MA term; *d* is the number of differencing required to make the time series stationary. ACF (autocorrelation) and PACF (partial autocorrelation) were adopted to determine the optimal parameters for models in the paper, expressed as ARIMA (2 1 2)for DO, ARIMA (2, 1, 2) for COD_{Mn}, ARIMA (5, 1, 5) for NH_3-N, and ARIMA (5, 1, 2) for TP.

SVR is a non-linear regression approach that is based on statistical learning theory (Smola and Schölkopf, 2004). The foundation of this strategy is to transfer the original input space into a new hyperspace using a non-linear transformation approach (kernel functions). Penalty parameter *C*, Kernel function and Kernel parameter *σ* are the main factors that affect the prediction accuracy of the SVR model. Radial Basis Function (RBF) may be the best performing algorithm for SVR models (Kooh et al., 2022). Penalty factors and kernel function parameters in the established SVR models were 2.4 and 8 for DO, 1.2 and 3 for COD_{Mn}, NH_3-N and TP, respectively.

RNN is a feedforward neural network that can store past information

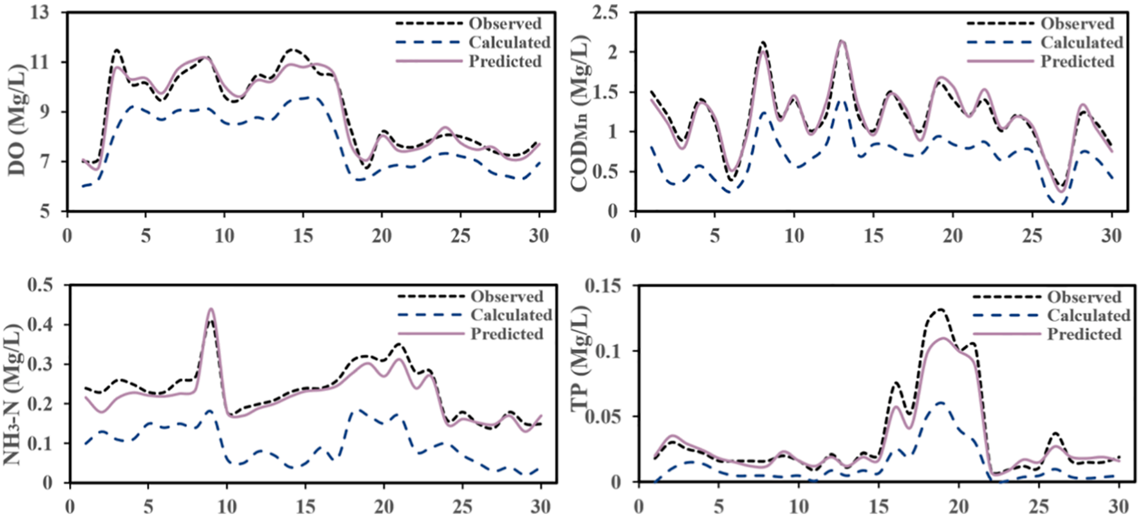


Fig. 8. The prediction performance of the SOD-VGG-LSTM model for extreme values.

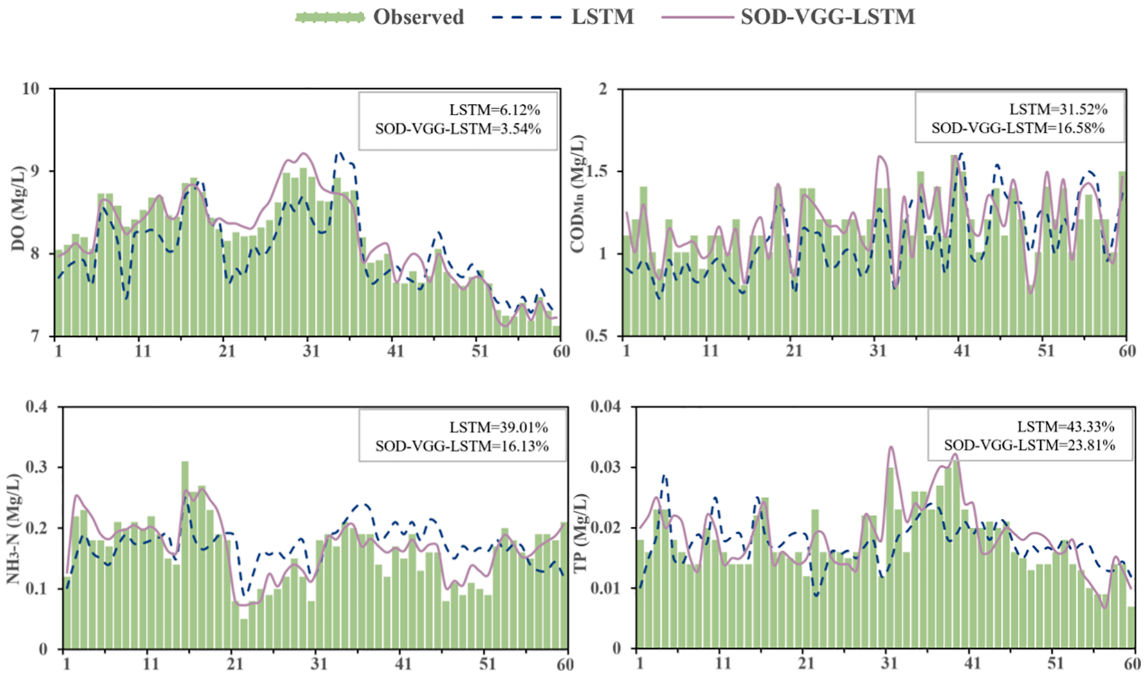


Fig. 9. Comparison of the predicted results between LSTM and SOD-VGG-LSTM.

in time series data by introducing state variables (Zhang et al., 2021b). The input of the hidden layer for RNN consists of two parts, including the output of the network layer at a given time and the output of the hidden layer at the previous time. The nodes of the hidden layer are connected to each other. This can ensured that the output of the current hidden layer can be affected by the output of the previous layers. Therefore, RNN model can deal with time series data well. The optimal structure of RNN adopted in this study had three hidden layers with 30 neurons in each layer.

The results of RMSE, MAE, and SMAPE for four models were presented in Table 3. According to the statistics, the proposed SOD-VGG-LSTM model achieved the best prediction results compared with other models. Among four state-of-the-art prediction models, ARIMA was the worst based on all statistics. Compared with ARIMA model, SVR model had a higher accuracy in water quality prediction, but the prediction accuracy was still lower. Neural networks for processing sequence data including RNN and SOD-VGG-LSTM could help in improving the model performance. Results in Table 3 reflected the stability and robustness of the established models.

The DO, COD_{Mn}, NH_3-N, and TP concentration predicted by different models including ARIMA, SVR, RNN, and SOD-VGG-LSTM were showed in Fig. 10. Among them, ARIMA and SVR models could reflect the time-varying trend of pollutants concentration in river, but neither model had some shortcomings in extreme value predict. This might be because ARIMA model was based on sliding average and autoregression. The predicted results of ARIMA model were close to historical average. ARIMA might be more appropriate when the true value did not fluctuate very strongly. SVR model failed to consider the impact of time series data on prediction results (the impact of the previous time on the next time), which might result in a reduction in the accuracy of extreme value predicting. In contrast, RNN and SOD-VGG-LSTM models achieved good performance in water quality prediction. SOD-VGG-LSTM model that coupled mechanistic model and spatial data had higher water quality prediction accuracy than RNN model. The evaluation results showed that SOD-VGG-LSTM achieved 3.2% – 39.3% higher R^2 than ARIMA, SVR and RNN. The established SOD-VGG-LSTM model in the paper could provide a new method for water quality prediction, especially for water quality prediction affected by NPS pollution.

Table 3
Prediction performance of four compared models.

Indicator	Method	RMSE	MAE	SMAPE(%)
DO	ARIMA	0.424	0.368	0.055
	SVR	0.350	0.319	0.032
	RNN	0.292	0.243	0.022
	SOD-VGG-LSTM	**0.261**	**0.225**	**0.014**
COD_{Mn}	ARIMA	0.252	0.204	0.180
	SVR	0.189	0.141	0.120
	RNN	0.140	0.091	0.081
	SOD-VGG-LSTM	**0.088**	**0.051**	**0.046**
NH_3-N	ARIMA	0.063	0.048	0.284
	SVR	0.056	0.032	0.188
	RNN	0.043	0.022	0.132
	SOD-VGG-LSTM	**0.017**	**0.012**	**0.078**
TP	ARIMA	0.021	0.016	0.621
	SVR	0.010	0.007	0.223
	RNN	0.007	0.005	0.155
	SOD-VGG-LSTM	**0.005**	**0.004**	**0.132**

6. Conclusions

A hybrid deep learning model coupling with SOD, VGG, and LSTM modular was developed in the paper. The training dataset was constructed by a set of time-series data, including hydrometeorological parameters, pollutant parameters, error sequence, and spatial feature sequences. The error sequence was calculated by the SOD modular, and the spatial feature was extracted by the VGG modular. The error sequence of pollutant concentration was used as output of LSTM for error and water quality prediction. The established model could not only overcome the problem of extreme value prediction, but also reflect the impact of spatial characteristics at different times or regions on water quality.

The performance of the established model was evaluated and compared with three state-of-the-art prediction models: ARIMA, SVR, and RNN models. The indicators of DO, COD_{Mn}, NH_3-N, and TP were selected as performance evaluation indicators, the RMSE index of the established model were 0.261, 0.088, 0.017, and 0.005, respectively. The results showed that the established SOD-VGG-LSTM model could predict the water quality change caused by NPS pollution well. SOD-

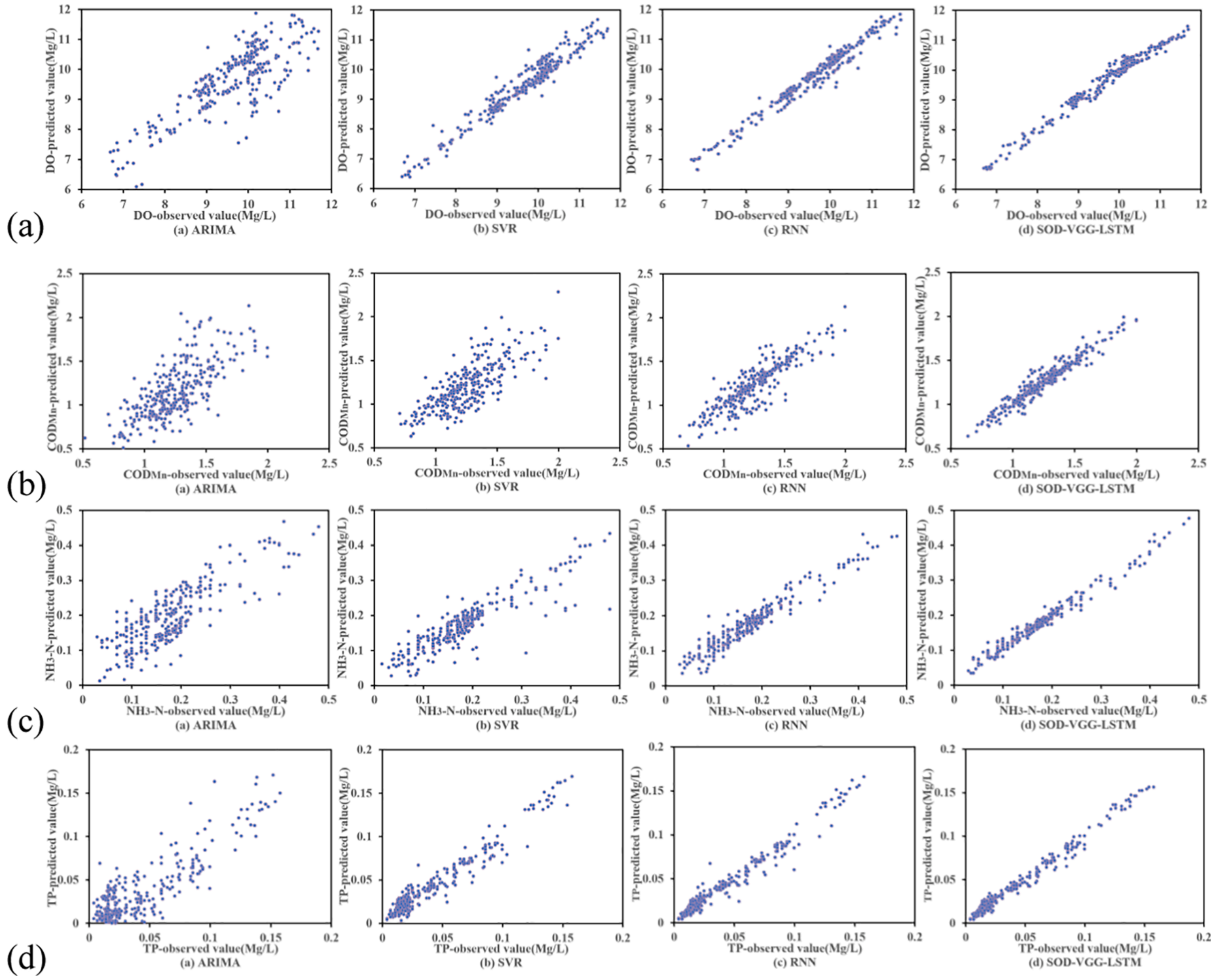


Fig. 10. Scatter plots of comparison models, (a) DO, (b) COD_{Mn}, (c) NH_3-N, (d) TP.

VGG-LSTM model had higher accuracy than ARIMA, SVR, and RNN model in water quality prediction. The evaluation results also showed that the established model could improve the prediction accuracy of the extreme value for water quality with coupling mechanism method and deep learning method.

The model framework proposed in this study can not only effectively solve the defect that traditional deep learning methods can not couple point and surface data, but also overcome the problem that mechanism models can not predict the changes of hydrology or water quality on the hourly or minute time scale. The model proposed in the study is an intelligent watershed water quality prediction model, which can be the key link to solve the non point source pollution in the development stage of intelligent water conservancy. Also, the proposed model can provide effective decision support for the control and risk management of the basin flood forecast in the future. Because the established model can provide timely and efficient early warning of the hydrological process. The model proposed in this study has great application potential.

CRediT authorship contribution statement

Hang Wan: Conceptualization, Methodology, Writing – original draft. **Rui Xu:** Validation, Formal analysis. **Meng Zhang:** Data curation, Software. **Yanpeng Cai:** Methodology, Formal analysis. **Jian Li:** Validation, Software. **Xia Shen:** Data curation, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (GML2019ZD0403), the National Natural Science Foundation of China (52009023); Guangxi Natural Science Foundation (2021JJA170096); Guangxi Key Research and Development Program (AB21196063); Guilin Science and Technology Development Program (20190213-1).

References

Baek, S.-S., Ligaray, M., Pyo, J., Park, J.-P., Kang, J.-H., Pachepsky, Y., Chun, J.A., Cho, K.H., 2020. A novel water quality module of the SWMM model for assessing Low Impact Development (LID) in urban watersheds. J. Hydrol. 586, 124886.

Bahaa, M.K., Ayman, G.A., Hussein, K., Ashraf, E., 2012. Application of artificial neural networks for the prediction of water quality variables in the Nile Delta. J. Water Resour. Prot. 4 (6), 388–394.

Bahman, M., Mohammad, H.N., Fereydoun, G., Sepehr, D., 2018. Assessing the impacts of climate change on the quantity and quality of agricultural runoff (Case Study: Golgol River Basin). Irrig. Drain. 67, 17–28.

Chen, Y.S., Jiang, H.L., Li, C.Y., Jia, X.P., Ghamisi, P., 2016. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 54 (10), 6232–6251.

Cui, Q., Wang, X., Li, C., Cai, Y., Liang, P., 2016. Improved thomas–fiering and wavelet neural network models for cumulative errors reduction in reservoir inflow forecast. J. Hydro-environ. Res. 13, 134–143.

Deng, Y., Zhou, X., Shen, J., Xiao, G.e., Hong, H., Lin, H., Wu, F., Liao, B.-Q., 2021. New methods based on back propagation(BP) and radial basis function(RBF) artificial neural networks(ANNs) for predicting the occurrence of haloketones in tap water. Sci. Total Environ. 772, 145534.

Dong, C., Huang, G., Cheng, G., Zhao, S., 2018. Water Resources and Farmland Management in the Songhua River Watershed under Interval and Fuzzy Uncertainties. Water Resour. Manage. 32 (13), 1–24.

Huang, N., Wang, H.Y., Lin, T., Liu, Q.M., Huang, Y.F., Li, J.X., 2016. Regulation framework of watershed landscape pattern for non-point source pollution control based on 'source-sink' theory: A case study in the watershed of Maluan Bay, Xiamen City, China. J. Appl. Ecol. 27 (10), 3325–3334.

Huang, W.R., Simon, F., 2002. Neural network modeling of salinity variation in Apalachicola River. Water Res. 36 (1), 356–362.

Hu, D., Zhang, C., Ma, B.o., Liu, Z., Yang, X., Yang, L., 2020. The characteristics of rainfall runoff pollution and its driving factors in Northwest semiarid region of china-A case study of Xi`an. Sci. Total Environ. 726, 138384.

Senent-Aparicio, J., Jimeno-Sáez, P., Bueno-Crespo, A., Pérez-Sánchez, J., Pulido-Velázquez, D., 2019. Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction. Biosyst. Eng. 177, 67–77.

Jiang, Y.Q., Li, C.L., Sun, L., Guo, D., Zhang, Y.T., Wang, W.H., 2021. A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. J. Cleaner Prod. 318 (8), 128533.

Kooh, M.R.R., Thotagamuge, R., Chou Chau, Y.-F., Mahadi, A.H., Lim, C.M., 2022. Machine learning approaches to predict adsorption capacity of Azolla pinnata in the removal of methylene blue. J. Taiwan Inst. Chem. Eng. 132, 104134.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90.

Li, C., Li, S.-L., Yue, F.-J., Liu, J., Zhong, J., Yan, Z.-F., Zhang, R.-C., Wang, Z.-J., Xu, S., 2019. Identification of sources and transformations of nitrate in the Xijiang River using nitrate isotopes and Bayesian model. Sci. Total Environ. 646, 801–810.

Liu, Z., Tong, S.T.Y., 2015. Using HSPF to model the hydrologic and water quality impacts of riparian land-use change in a small watershed. J. Environ. Inform. 17 (1), 15–24.

Mcilwaine, B., Rivas, C.M., 2020. JellyNet: The convolutional neural network jellyfish bloom detector. Int. J. Appl. Earth Obs. Geoinf. 97, 102279.

Navideh, N., Latif, K., Sabahattin, I., 2020. Water quality prediction using SWAT-ANN coupled approach. J. Hydrol. 590, 125220.

Nitzan, F., Efrat, K., Hadas, M., Yuval, S., 2021. Prediction of wastewater treatment quality using LSTM neural network. Environ. Technol. Innovation 23, 101632.

Paparrizos, S., Maris, F., 2017. Hydrological simulation of Sperchios river basin in central Greece using the MIKE SHE model and geographic information systems. Appl. Water Sci. 7 (2), 591–599.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651.

Simonyan, K., Zisserman, A., 2004. Very deep convolutional networks for large-scale image recognition. In: The 3rd International Conference on Learning Representations, San Diego, Canada.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Tiyasha, Tung, T.M., Yaseen, Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 585, 124670.

Wan, H., Mao, Y., Cai, Y., Li, R., Feng, J., Yang, H., 2021a. An SPH-based mass transfer model for simulating hydraulic characteristics and mass transfer process of dammed rivers. Eng. Comput. https://doi.org/10.1007/s00366-021-01354-2.

Wan, H., Tan, Q., Li, R., Cai, Y., Shen, X., Yang, Z., Shen, X., 2021b. Incorporating fish tolerance to supersaturated total dissolved gas for generating flood pulse discharge patterns based on a simulation optimization approach. Water Resour. Res., 57, e2021WR030167.

Wang, F., Wang, X., Chen, B., Zhao, Y., Yang, Z., 2013. Chlorophyll a Simulation in a Lake Ecosystem Using a Model with Wavelet Analysis and Artificial Neural Network. Environ. Manage. 51 (5), 1044–1054.

Wijesiri, B., Egodawatta, P., McGree, J., Goonetilleke, A., 2015. Influence of pollutant build-up on variability in wash-off from urban road surfaces. Sci. Total Environ. 527–528, 334–350.

Xie, Y.L., Xia, D.X., Ji, L., Huang, G.H., 2018. An inexact stochastic-fuzzy optimization model for agricultural water allocation and land resources utilization management under considering effective rainfall. Ecol. Ind. 92, 301–311.

Xu, R., Deng, X., Wan, H., Cai, Y., Pan, X., 2021. A deep learning method to repair atmospheric environmental quality data based on Gaussian diffusion. J. Cleaner Prod. 308, 127446.

Xu, Y., Hu, C., Wu, Q., Jian, S., Li, Z., Chen, Y., Zhang, G., Zhang, Z., Wang, S., 2022. Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation. J. Hydrol. 608, 127553.

Xu, Z.F., Li, J.C., Mo, D.Q., et al., 2010. Study on the Present State of Rural Water Pollution in Li River Valley and Comprehensive Treatment. Environ. Sci. Technol. 33 (12F), 644–650.

Yan, J., Chen, X., Yu, Y., Zhang, X., 2019. Application of a parallel particle swarm optimization-long short term memory model to improve water quality data. Water 11 (7), 1317.

Zhang, Q., Li, Z., Zhu, L.u., Zhang, F., Sekerinski, E., Han, J.-C., Zhou, Y., 2021a. Real-time prediction of river chloride concentration using ensemble learning. Environ. Pollut. 291, 118116.

Zhang, X., Liu, L.u., Long, G., Jiang, J., Liu, S., 2021b. Episodic memory governs choices: an RNN-based reinforcement learning model for decision-making task. Neural Network 134, 1–10.

Zhou, W., Zhu, Z., Xie, Y., Cai, Y., 2021. Impacts of rainfall spatial and temporal variabilities on runoff quality and quantity at the watershed scale. J. Hydrol. 603, 127057.

Zuo, Q., Wu, Q., Yu, L., Li, Y., Fan, Y., 2021. Optimization of uncertain agricultural management considering the framework of water, energy and food. Agric. Water Manag. 253 (1), 106907.