MDPI

Article

Feature Extraction and Prediction of Water Quality Based on Candlestick Theory and Deep Learning Methods

Rui Xu [1], Wenjie Wu [1], Yanpeng Cai [2,3], Hang Wan [2,3,*], Jian Li [1], Qin Zhu [2,3] and Shiming Shen [1]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[2] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China
[3] 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
* Correspondence: wanhang@gdut.edu.cn

Abstract: In environmental hydrodynamics, a research topic that has gained popularity is the transmission and diffusion of water pollutants. Various types of change processes in hydrological and water quality are directly related to meteorological changes. If these changing characteristics are classified effectively, this will be conducive to the application of deep learning theory in water pollution simulation. When periodically monitoring water quality, data were represented with a candlestick chart, and different classification features were displayed. The water quality data from the research area from 2012 to 2019 generated 24 classification results in line with the physics laws. Therefore, a deep learning water pollution prediction method was proposed to classify the changing process of pollution to improve the prediction accuracy of water quality, based on candlestick theory, visual geometry group, and gate recurrent unit (CT-VGG-GRU). In this method, after the periodic changes of water quality were represented by candlestick graphically, the features were extracted by the VGG network based on its advantages in graphic feature extraction. Then, this feature and other scenario parameters were fused as the input of the time series network model, and the pollutant concentration sequence at the predicted station constituted the output of the model. Finally, a hybrid model combining graphical and time series features was formed, and this model used continuous time series data from multiple stations on the Lijiang River watershed to train and validate the model. Experimental results indicated that, compared with other comparison models, such as the back propagation neural network (BPNN), support vector regression (SVR), GRU, and VGG-GRU, the proposed model had the highest prediction accuracy, especially for the prediction of extreme values. Additionally, the change trend of water pollution was closer to the real situation, which indicated that the process change information of water pollution could be fully extracted by the CT-VGG-GRU model based on candlestick theory. For the water quality indicators DO, COD_{Mn}, and NH_3-N, the mean absolute errors (MAE) were 0.284, 0.113, and 0.014, the root mean square errors (RMSE) were 0.315, 0.122, and 0.016, and the symmetric mean absolute percentage errors (SMAPE) were 0.022, 0.108, and 0.127, respectively. The established CT-VGG-GRU model achieved superior computational performance. Using the proposed model, the classification information of the river pollution process could be obtained effectively and the time series information could also be retained, which made the application of the deep learning model to the transmission and diffusion process of river water pollution more explanatory. The proposed model can provide a new method for water quality prediction.

Keywords: candlestick theory; deep learning; water quality prediction

1. Introduction

The simulation and prediction of water quality are difficult, as they are affected by various factors, such as complex transport process and climate change [1,2]. Water quality is

obviously related to meteorological conditions in the rain-sourced rivers, and there are various changes in the process of water pollution. The Soil and Water Assessment Tool (SWAT) is the most widely used physical model, but with some limitations, including difficulty in parameter calibration and complex model construction [3]. Additionally, physical models typically require high levels of expertise to implement. The process classification of water pollution has not been fully considered in previous artificial intelligence models, and the focus of these models was only on the statistical law of the numerical sequence between input and output. The change process of water pollution could not be effectively classified, and the prediction accuracy of the model was affected. Therefore, solving this problem would be useful for the application of artificial intelligence models to water pollution simulations.

Traditional physical models, such as the SWAT model, have been used to simulate hydrological and pollution processes and the transport of runoff, sediment, and nutrients in the river, but problems of redundant parameters and complex structure are difficult to ignore [4]. The MIKE SHE model had been used to simulate the main hydrological processes in the water cycle, including flow movement, water quality, and sediment transport. This model requires the acquisition of high-resolution data on the physical properties of the watershed, which was time-consuming and difficult to process [5]. The water quality analysis simulation program (WASP) model had been used to simulate the migration and transformation of conventional pollutants in water, including dissolved oxygen, biological oxygen consumption, nutrients, and algal pollution, but there were limitations in simulating the eutrophication process [6]. These outcomes demonstrated that the traditional models were applied primarily to the large-scale macro-analysis and had good practicability. Meanwhile, it was necessary to collect information on hydrology, meteorology, geology, land use, farming methods, crop types, and regional economy with a huge quantity of data. Some areas, however, did not have sufficient basic data to support the simulation of traditional physical models. It was difficult to calibrate the parameters, especially in the accurate simulation and prediction of hydrology and water quality.

At present, in the previously low-developed areas lacking data, more monitoring stations have been established along the key rivers, and the image data can be obtained through remote sensing at a low cost. Therefore, some scholars tried to use machine learning models and deep learning models to study water quality simulation and prediction. Machine learning models, such as the backpropagation neural network (BPNN), had been used to predict water quality parameters, including chlorophyll-a (Chl-a), dissolved oxygen, and biochemical oxygen demand, but they were slow to converge and easy to reach extreme minimum value [7,8]. The SVR model had been used to predict water quality indices, such as ammonia nitrogen, dissolved oxygen, and chemical oxygen demand, but was hard to implement on large-scale training samples [9–11]. The ANN models had been used to predict river salinity, dissolved oxygen, and chlorophyll-a, but the model could not learn the state characteristics between time series water quality data [12]. In general, these machine learning models were based on statistical learning methods, which were not integrated with the physical laws and thus could not achieve accurate feature recognition and extraction in a large quantity of data. Therefore, it was suggested that these models were not suitable for the study of time series information of water quality data.

In recent years, deep learning models with the characteristics of multilayer feedback simulation have become mainstream in water quality simulation, overcoming the critical timing limitations of machine learning models [13]. Affected by various environmental parameters, water pollution data had included continuous nonlinear time series data with obvious correlation among those cycles, which had been consistent with the structure of the time series deep learning model [14]. For example, the recurrent neural network (RNN) model had been used to predict water quality indicators, including total phosphorus, total nitrogen, dissolved oxygen, and ammonia nitrogen, but it could not solve the problems of long-distance dependence and gradient explosion [15,16]. The long short-term memory (LSTM) model had been used to predict dissolved oxygen and water temperature changes in Taihu Lake and Victoria Harbor in China, but there were disadvantages of complex

structure and long model training period [17]. The convolutional neural network–LSTM (CNN-LSTM) model had been used to predict dissolved oxygen and chlorophyll-a [18,19]. Among these deep learning models, CNN is more suitable for classification simulation, and RNN and LSTM are more suitable for time series prediction [20]. For these deep learning models, however, continuous sequential data were used directly only as the input of the model. In the black-box model established by the connection between input and output, the classification features of water pollutant transport and diffusion were not sufficient to be extracted, and the corresponding relationship between the network structure and the water pollution process had not been explained. Therefore, these studies on water quality simulation could not reflect the physical law of water quality variable transmission, and the prediction accuracy was not high enough.

Different classification characteristics were found in the process of water pollution transmission and diffusion, and the change process of water pollution transmission was reflected. The candlestick theory was originally used for stock price changes over time [21,22]. Its graphical advantages could accurately reflect the change process of stocks in a period of time [23,24]. Many scholars had used the candlestick chart to extract the change characteristics of time series data for trend prediction [25,26]. For example, deep learning models based on candlestick theory were used to analyze patterns within candlestick charts and predict future movements of the stock market [27]. A method based on the candlestick chart was proposed to predict the $PM_{2.5}$ concentration in Guilin, China, that used the trend in the candlestick chart to reflect the changing law of air pollutants [28]. Water quality data are also typically time series data, so continuous sequential water quality data can be represented in graphical form in the candlestick chart, which corresponds to different types of classification. We studied whether the water quality data in graphical form had classification rules. The classification features could be extracted by the convolutional neural network and used for the classification of different trends in water pollution. Then, combined with the time series deep learning model, candlestick theory, visual geometry group, and gate recurrent unit (CT-VGG-GRU), a hybrid model for classification and prediction was constructed. Based on the proposed model, the water pollution classification process information was extracted and completely retained, which showed that each classification type could be predicted in a time series. Therefore, this meant that the CT-VGG-GRU hybrid model was a suitable combination for the application of deep learning theory to the simulation of water pollution.

The rest of this article is organized as follows. Section 2 explains the theoretical knowledge of the candlestick theory and the theoretical basis for the candlestick theory to be applied to the water pollution diffusion. Additionally, this section discusses the geographical advantages of the research area. Section 3 explains the overall framework structure of this study, including data acquisition and preprocessing, candlestick chart generator, VGG feature extraction, and GRU time series forecasting. Section 4 provides the experimental results, comparing this study with other models. Finally, the work of this paper is summarized and future research directions are prospected in Section 5.

2. Problem Scenario

2.1. Candlestick Graphical Representation of Pollution Process Classification

On the basis of the concentration-period change trend, the time-sequence characteristics of water pollution diffusion were extracted to classify the change trends. In the financial field, candlestick theory had been widely used to extract time series features [29]. Therefore, the candlestick theory could be introduced into the water pollution diffusion model. Various classification types were reflected in graphical form, different changing types were classified and processed, and the prediction accuracy of the model was improved.

The candlestick theory, commonly known as the "K-line", was developed by the Japanese author Munehisa Homma in the 18th century to record the fluctuation of rice prices within the cycle. In 1991, candlestick theory was used by Steve Nison to record stock price changes over time [30]. The candlestick chart is composed of the opening

price, the lowest price, the highest price, and the closing price, and its color represents the relationship between the opening price and the closing price. Water quality data are also recorded and counted by hours, days, months, and years. The periodic changes in water quality data were represented by candlestick charts and the key value of the periodic original data was reflected. The fluctuation relationship of the candlestick chart also corresponded to the process of water quality change, which contained the mechanism of hydraulics and water quality diffusion. According to the research results of Hu et al., 103 comprehensive classification specifications were proposed, including 31 three-day candlestick chart classification combinations [31]. Based on the water quality data of the prediction station from 2012 to 2019, 24 corresponding candlestick chart classification combinations were identified, which covered various classification combinations, as shown in Table 1.

Table 1. Twenty-four categories of the three-day candlestick chart combinations for the water quality data at the prediction station.

Species	1	2	3	4	5	6	7	8
Candlestick chart								
Species	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Candlestick chart								
Species	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Candlestick chart								

In the securities industry, the factors that affect stock prices are too complex and have not yet been supported by clear mathematical theories, so the application and classification of candlestick theory are still in the stages of perceptual knowledge. However, the candlestick chart could be applied to simulate the change process of water quality, which had a direct and definite relationship with the hydraulic process. As shown in the candlestick combination category 1 in Table 1, while the pollutant concentration dropped significantly on the first day, it had no change on the second day, and it increased significantly on the third day. The pollutant concentration for the next day could be predicted based on such a change trend to simulate the diffusion process of water pollution. Therefore, there was a theoretical basis that the candlestick chart can be applied to the water pollution diffusion.

2.2. Correspondence between Candlestick Characteristics and the Physical Model of Water Quality Diffusion

On the basis of the physical mechanism of water quality diffusion, each candlestick chart classification was scientifically explained, which laid the theoretical foundation for the subsequent deep learning theory applied to water pollution simulation. The water quality diffusion physical models were used to describe the migration and transformation laws of various pollutants [32]. According to the migration and transformation process of pollutants in water and the mass balance theory, the mathematical model of pollutant migration and transformation was established to predict the concentration of pollutants at the target station by simulating the diffusion process. The physical model of water quality diffusion is shown in Equation (1) [33]:

$$c(x,y) = \exp\left(-K\frac{x}{86400\cdot u}\right)\left\{C_0 + \frac{C_pQ_p}{H(\pi M_y x u)^{1/2}}\left[\exp\left(-\frac{uy^2}{4M_y x}\right) + \exp\left(-\frac{u(2B-y)^2}{4M_y x}\right)\right]\right\} \tag{1}$$

where $c(x, y)$ is the concentration of pollutants at the target station; K is the degradation coefficient; x is the distance between the target station and the monitoring station; y is the transverse distance between the target station and the monitoring station; u is the longitudinal average velocity of the river; C_0 is the concentration of pollutants at the monitoring station and is equivalent to the source concentration; C_p is the concentration of wastewater discharge at the monitoring station; Q_p is the amount of wastewater discharged from the monitoring station; H is the average depth of the river; B is the average width of the river; and M_y is the transverse dispersion coefficient. According to the source concentration C_0, the river longitudinal average velocity u, and the degradation coefficient K, the physical model of water quality diffusion was analyzed. The changes for each of the influencing factors were closely related to the graphical form of the candlestick chart, whereas other factors were constant or had little influence.

The influencing factors of the water diffusion physical model correspond to the basic types of candlestick charts, which laid a theoretical foundation for the candlestick charts to be applied to water pollution simulation. The water quality data were converted into candlestick charts based on the candlestick chart generator. According to the initial, minimum, maximum, and end values of water quality data, nine basic forms of candlestick charts were determined. After analysis, a corresponding relationship between the basic form and the influencing factors was obtained, as shown in Table 2.

Table 2. Correspondence between nine basic candlestick chart forms and water quality diffusion factors.

Species	1	2	3	4	5	6	7	8	9
Candlestick chart									
C_0 increase	Y	Y	Y	Y	Y	N	N	N	N
C_0 decrease	N	N	N	N	N	Y	Y	Y	Y
u increase	Y	Y	Y	Y	N	N	N	N	N
u decrease	N	N	N	N	Y	Y	Y	Y	Y
K increase	N	N	N	N	N	Y	Y	Y	Y
K decrease	Y	Y	Y	Y	Y	N	N	N	N

The white candlestick chart indicates that the ending value was higher than the initial value, meaning that the pollutant concentration was rising. The black candlestick chart indicates that the end value was lower than the initial value, meaning that the pollutant concentration was falling. The Doji candlestick chart indicates that the end value was equal to the start value, which means that the pollutant concentration was not changed on the same day. *Y* means the change and *N* means no change. In basic form 1, the pollutant concentration increased significantly, C_0 and u increased, and K decreased. In basic form 6, the concentration of pollutants decreased significantly, C_0 and u decreased, and K increased. The basic form, similar to the candlestick chart, had corresponded to the changes of various influencing factors. On the basis of the physical mechanism and mathematical theory, the Visual Geometry Group (VGG) model, which was an improved convolutional neural network model, was used to extract the features in the candlestick chart. Therefore, the classification features of the water pollution process were retained.

2.3. Research Area

The Lijiang River watershed in Guilin is a typical karst landform river watershed belonging to the rain-sourced river and subtropical humid monsoon climate zone. In the watershed, there was concentrated rainfall, rising in summer and drying in winter, and a significant difference between precipitation in the wet period and the dry period. The maximum rainfall was up to 168 mm, and the flow and velocity changed significantly with rainfall [34]. The flow of the main river channel had obvious seasonal characteristics,

and the pollutant data had obvious classification characteristics, corresponding to various candlestick chart classification combinations. Therefore, the water quality diffusion equation could be fully reflected in the Lijiang River watershed, and it was suitable to verify the model designed in this study. The water quality data of the monitoring stations in the upstream areas were used as the research objects. Additionally, three water quality indicators of Yangshuo Station were the prediction object, including dissolved oxygen (DO), chemical oxygen demand (COD_{Mn}), and ammonia nitrogen (NH_3-N). The locations of monitoring stations in the Lijiang River watershed are shown in Figure 1.
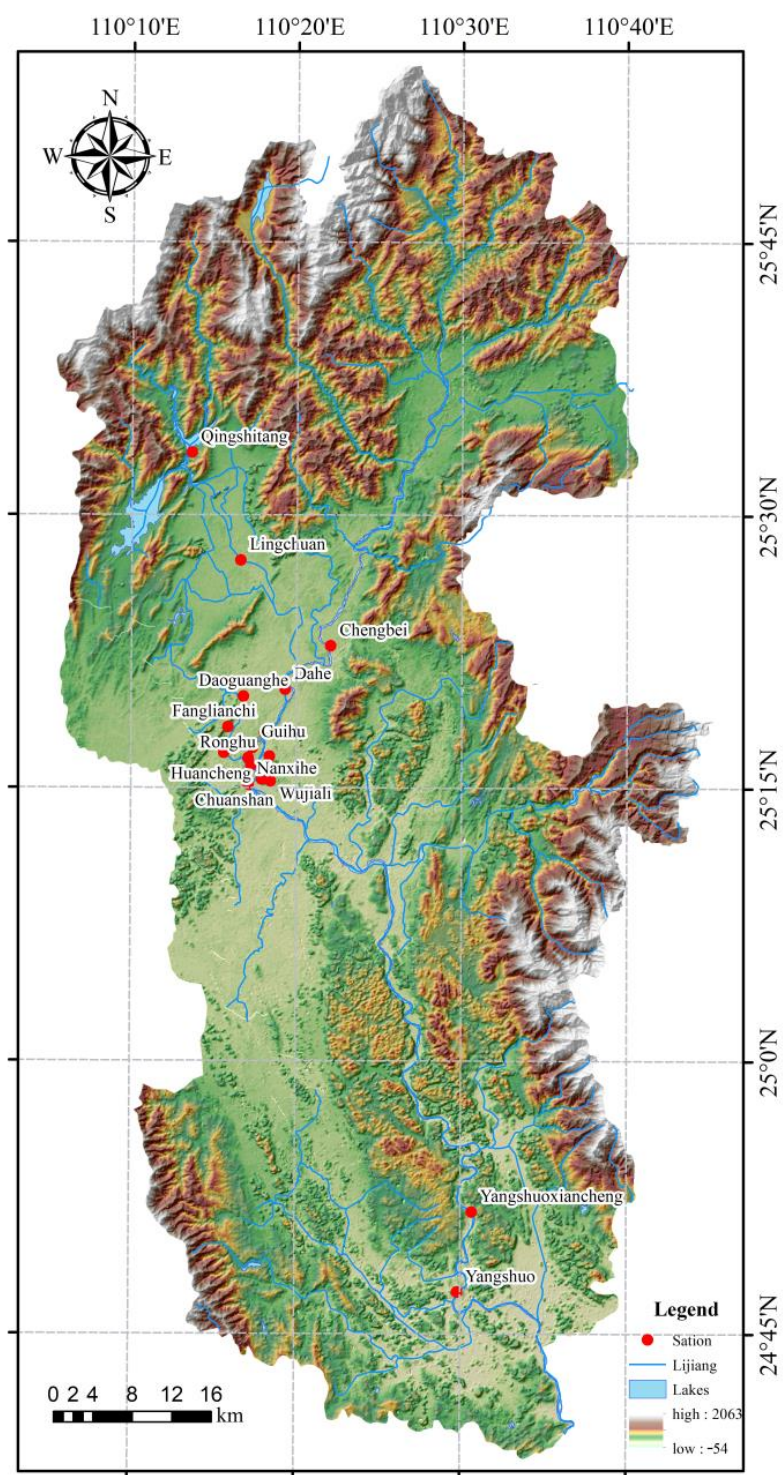

Figure 1. Distribution of stations in the Lijiang River Basin.

3. Methods

3.1. Framework

A water quality prediction framework based on candlestick chart classification is shown in Figure 2. On the basis of the theoretical connection between candlestick chart classification and water quality transmission diffusion mechanism, a CT-VGG-GRU model was designed by using the characteristics of VGG feature extraction and GRU timing prediction. It included data collection and preprocessing, candlestick chart generator, VGG feature extraction, GRU time series prediction, and results analysis. First, research data were collected from the online water quality monitoring station of the Lijiang River watershed in Guilin, including hydrometeorological parameters and pollutant parameters. Second, a candlestick chart generator was designed to generate three-day candlestick charts of the target station water quality data such as DO, COD_{Mn}, and NH_3-N. Third, the classification information of the water pollution process in the candlestick chart was extracted by VGG, and the water pollution process information was relatively completely preserved. Then, the extracted features were fused with other scene parameters of each station as the input of the time series deep learning model, GRU. Using the advantages of GRU in processing the time series data, the concentration of pollutants in the river was accurately predicted. The water quality data of the previous $n - 1$ days were used to predict the daily mean value of the water quality data of the nth day. In this paper, n was 4 and the sliding step was one day. Finally, MAE, RMSE, and SMAPE were used as evaluation parameters. The parameters were optimized continuously throughout the experimental process, and the effects were verified through the comparison of various models.
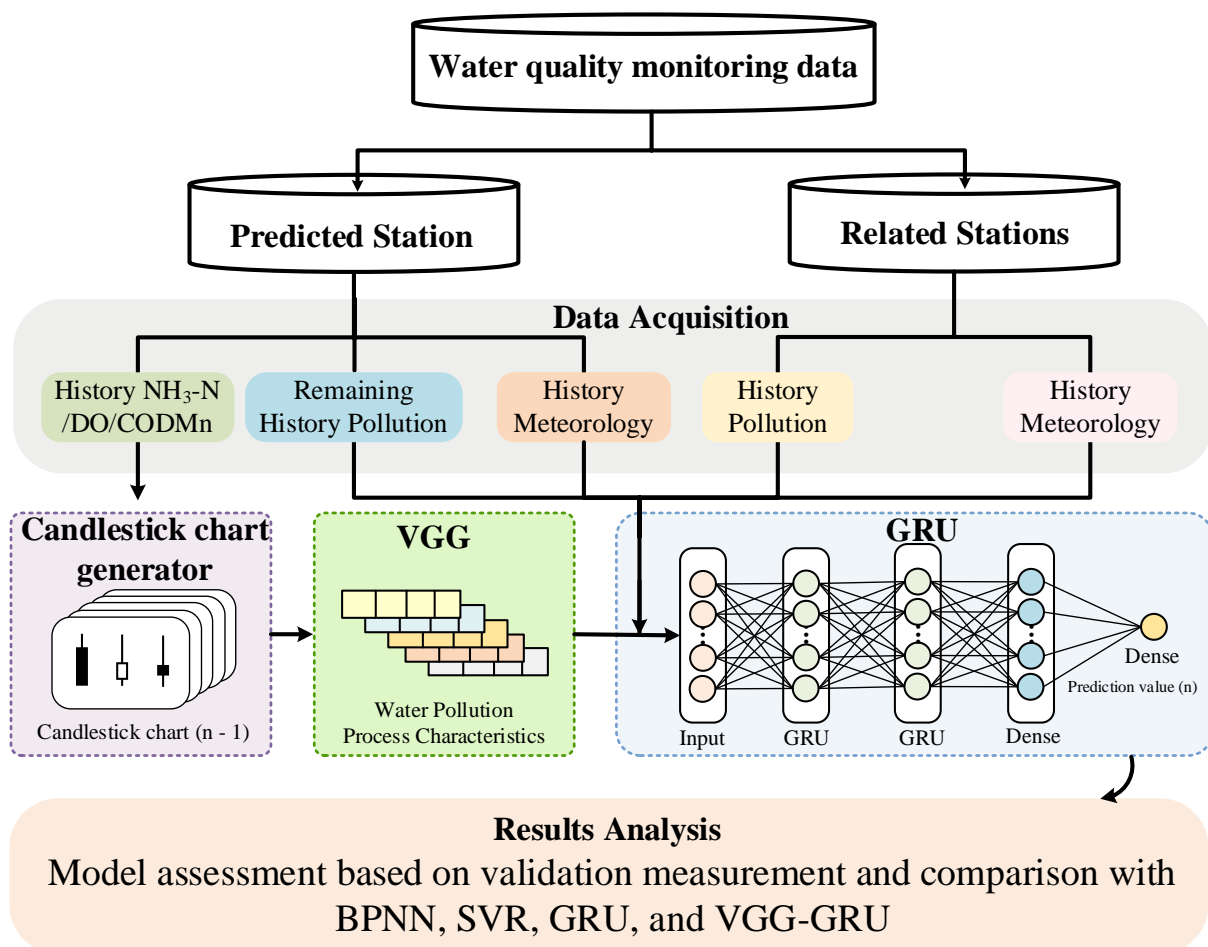

Figure 2. Framework of the proposed model.

3.2. Data Collection and Preprocessing

There are 47 online water quality monitoring stations in the Lijiang River watershed, and the common data are hourly value data. In Table 3, the basic dataset includes pollutant data and hydrometeorological data. Pollutant data include total phosphorus (TP), total nitrogen (TN), chemical oxygen demand (COD_{Mn}), ammonia nitrogen (NH_3-N), and dissolved oxygen (DO). Hydrometeorological data include conductivity (EC), pH, turbidity (TB), flow rate (Q), water temperature (WT), and rainfall (PCP). The core indices of water quality evaluation, DO, COD_{Mn}, and NH_3-N, were selected as the model prediction indicators, and the real situation in the water quality environment was objectively reflected.

Table 3. Pollution and hydrometeorology data parameters.

Data Category	Parameter	Unit
Water quality	NH_3-N	mg/L
	TP	mg/L
	TN	mg/L
	COD_{Mn}	mg/L
	DO	mg/L
Hydrometeorology	EC	μs/cm
	PH	Dimensionless
	TB	NTC
	Q	m^3/s
	WT	°C
	PCP	Mm

The data collected in this study were sorted based on time series. For continuous time series water quality data, the most important feature is the temporal correlation between the data, which contained many hidden laws and information. Because of equipment abnormality, network breakdown, and extreme weather, the record of water quality data at the monitoring station was abnormal. Data exception handling included three situations: deleting the negative value of obvious errors and missing data throughout the day; filling the local missing data with average value, and using the 3σ principle to analyze and eliminate outliers. To eliminate the dimensional influence between different indices and facilitate comprehensive analysis, the Z-score standardization method was selected to normalize the original monitoring data. The formula is shown in Equation (2):

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

where z is the normalized monitoring data, x is the original monitoring data, μ is the mean value of the original monitoring data, and σ is the standard deviation of the original monitoring data.

3.3. Design Principle of the Candlestick Chart Generator

The candlestick chart generator was designed to reflect the data changes of water quality during the period, and its principle is shown in Figure 3. The basic period was set to 24 hours, and the candlestick chart was generated by the initial, lowest, highest, and end values of the basic period. The continuous characteristics in the process of water quality change were reflected, which was based on the convolution idea of the CNN. Only time was convolved sequentially, and a day was set as a time unit. The water quality data formed a candlestick graph every three days by setting the sliding window size to three days, and the sliding step to one day. Therefore, the data of the 123, 234, 345, . . . , N – 1, N, and N + 1 days were used as the candlestick chart combination, corresponding to the continuous pollution process of water quality.
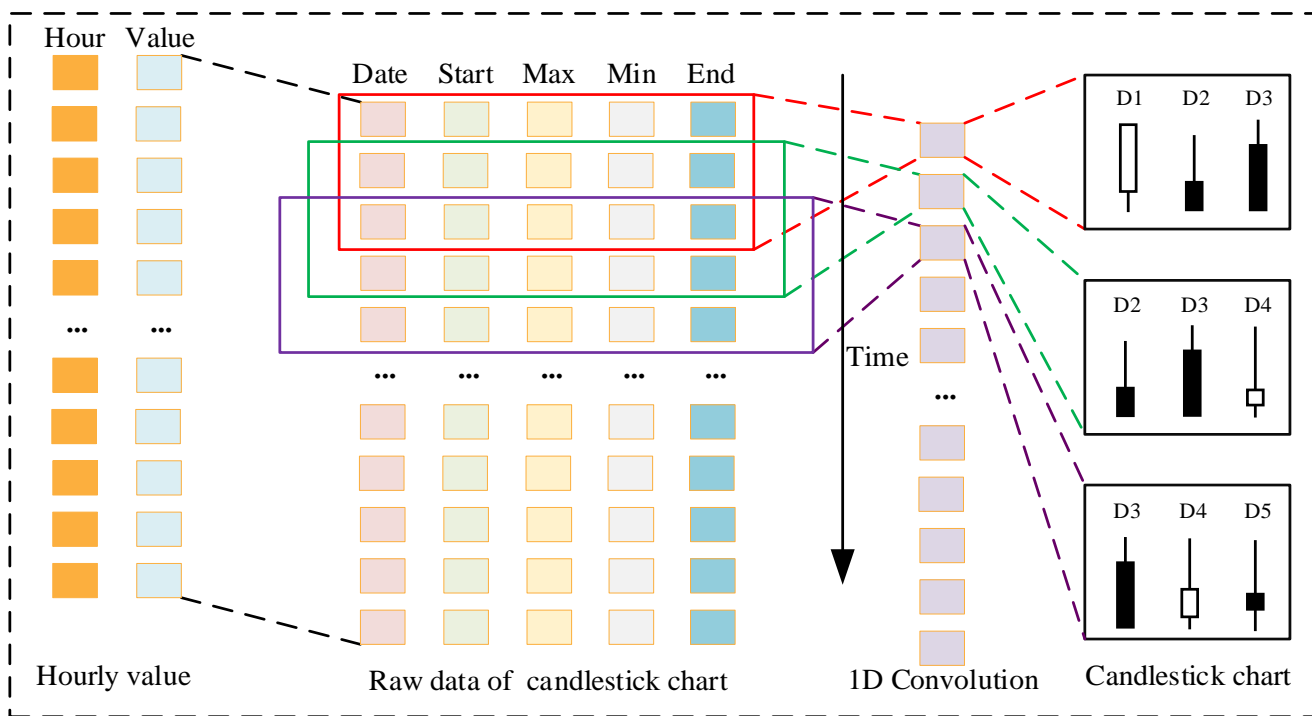
Figure 3. Principle of the candlestick chart generator.

3.4. Feature Extraction of Water Pollution Process Characteristics through VGG

The VGG network model, proposed by the Oxford Visual Geometry Group, is adapted from the CNN mode, which can be widely used to extract image features and save key information [35]. The characteristics of the candlestick chart extracted by VGG are shown in Figure 4. The local features of water pollution were extracted by the convolutional layer, which corresponded to the pollution transfer process. The pollution on the first day would affect the pollution on the second and third days, and such features were extracted by the convolutional layer. The extracted feature was further strengthened by the pooling layer, where only strong features were retained and weak features were discarded, which improved the accuracy of the pollution process classification. The global classification information of water pollution was integrated by the fully connected layer, which was equivalent to the pollution process classifier [36].
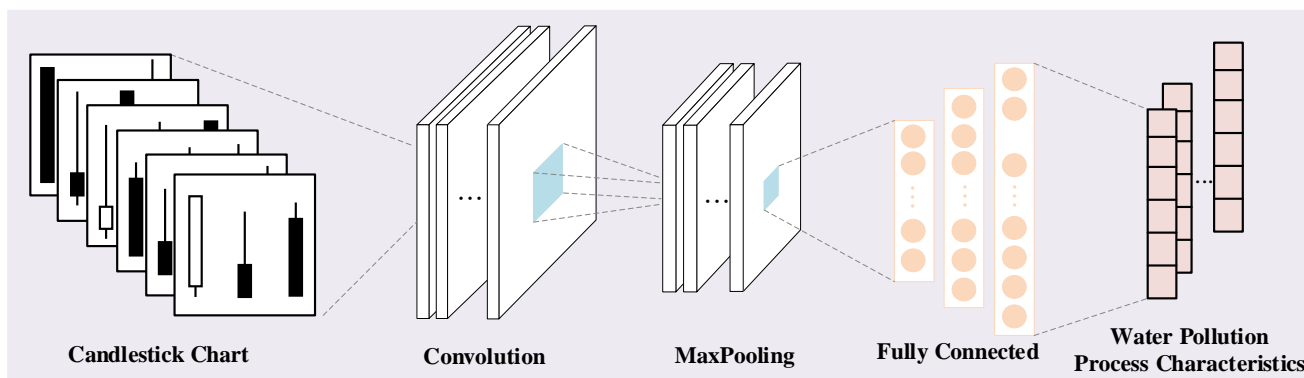

Figure 4. VGG extracted the candlestick chart features for classification.

As shown in Figure 4, the candlestick chart of water quality data generated by the candlestick chart generator for three consecutive days was used as the input of the VGG network. Corresponding to category 2 in Table 1, concentrations increased significantly

on the first day, decreased slightly on the second day, and decreased significantly on the third day. The variation in concentration on each day was extracted by shallow convolutional layers. The features extracted by the shallow convolutional layer, however, were integrated with the deep convolutional layer to obtain the characteristics of the overall water pollution process—that is, the change of the overall concentration in three days was equivalent to the characteristics of the water pollution transmission process being extracted. The information of the deep convolutional layer was integrated by the fully connected layer to obtain the classification result. Similar pollution processes were reflected in the candlestick chart. The VGG network structure had a unique fine-grained feature extraction method for images, which was able to classify pollution processes efficiently and retain the characteristics of pollution processes [37].

3.5. Time Series Prediction of Water Quality Data through GRU

The characteristic information of the water pollution process was extracted by VGG, combined with other scenario parameters of each station for time series prediction. The GRU network, which is a time series model improved and optimized based on LSTM, was adopted as the time series part of the model. Its structure is shown in Figure 5. Compared with LSTM, GRU has two gate structures: a reset gate and an update gate. The reset gate was used to control the degree of ignoring the water quality information at the previous moment—that is, the influence degree to which the water quality information at the previous moment on the current moment reflected the transmission process of water pollution. The update gate was used to control the degree to which the water quality information for the previous time was retained to the current state. The information for the previous time step and the current time step would affect the future water quality, which fully reflected the advantages of GRU in processing water quality time series data [38]. A special feature of these two gating mechanisms was that the water quality information in the long-term sequence could be saved and, thus, it would not be cleared over time or removed, making it irrelevant to the prediction [39].
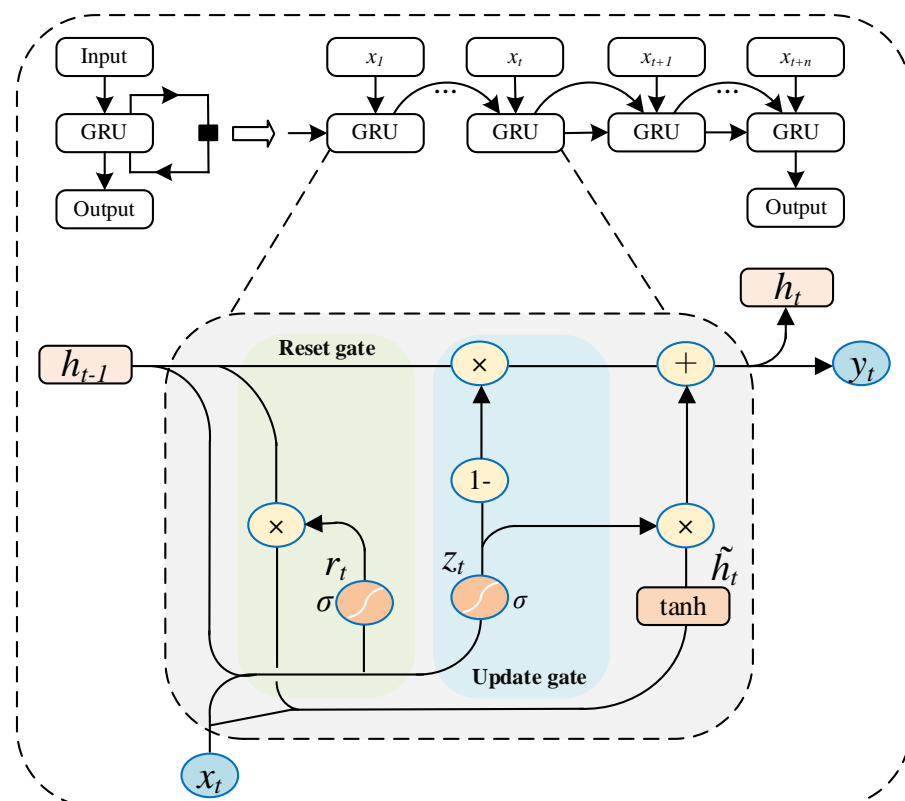

Figure 5. Architecture of the GRU.

In the GRU architecture of Figure 5, x_t was the combination of the classification information of the target station at the current moment and the water quality information of the relevant station; h_{t-1} was the water quality information retained at the previous moment; r_t was the output of the reset gate at the current moment, indicating the degree to which the water quality information at the previous moment will be retained—that is, the degree of influence on the water quality information at the current moment; z_t was the output of the update gate at the current moment, indicating the extent to which the water quality information at the previous moment was retained to the current moment—that is, the influence degree of the water quality information at the previous moment and the current moment on the next moment; and h_t was the overall output of the GRU at the current moment, which was retained as the water quality information at the current moment and continued to affect the water quality information at the next moment. This architecture explained why GRU could process continuous time series data. The specific calculation process is shown in Equations (3)–(9):

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{3}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{4}$$

$$\widetilde{h} = \tan h\left(W_{\widetilde{h}} \cdot [r_t * h_{t-1}, x_t]\right) \tag{5}$$

$$\widetilde{h} = \tan h\left(W_{\widetilde{h}} \cdot [r_t * h_{t-1}, x_t]\right) \tag{6}$$

$$y_t = \sigma(W_o \cdot h_t) \tag{7}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{9}$$

where [] denotes the connection of two vectors, $*$ denotes a matrix product, and σ denotes activation function.

The LSTM model contained three gate structures that could retain water quality information at the previous moment. If the data sequence was too long, it could avoid losing some water quality information during the training process. However, GRU had fewer gate structures and parameters, and it was not easy to overfit. This not only solved the problem of water quality information loss but also saved time and memory, so it was more efficient.

3.6. The CT-VGG-GRU Model for Water Quality Prediction

The chronological change law of the water quality data at the target station was reflected by the candlestick theory. The classification features in the candlestick chart could be fully and efficiently extracted by VGG. Because of the long-term learning dependency of GRU, the long-term correlation of water quality data was able to be captured, and the periodic pattern of water quality data over a long period could be accessed. Therefore, based on the candlestick theory, VGG, and GRU, the CT-VGG-GRU hybrid model was used to simulate the continuous transmission process of water pollution. Its structure is shown in Figure 6. First, the VGG model was used to filter the classification features of the candlestick chart in advance, reflecting the water pollution transmission process. Then, the extracted features were combined with the scenario parameters of each station, which together served as the input of the GRU model. Finally, the GRU model was used to reflect the continuous water pollution process and achieve accurate predictions of the water quality data.
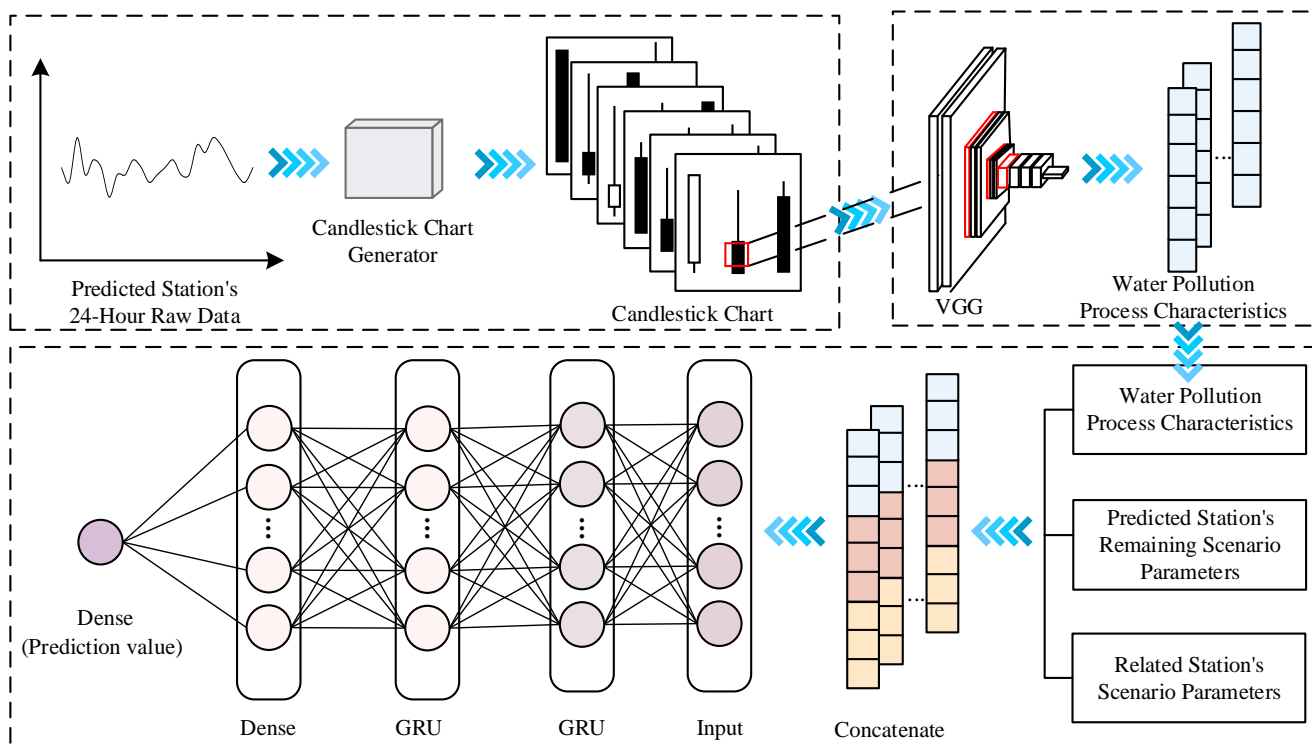
Figure 6. Architecture of the CT-VGG-GRU model.

4. Experiment Results and Analysis

4.1. Evaluation Criteria

The prediction performance of the proposed CT-VGG-GRU model was evaluated. First, the hyperparameters of the model were determined. Second, the performance of the CT-VGG-GRU model and the ordinary deep learning model VGG-GRU that did not consider the classification of the water pollution process were compared. Finally, compared with other models under the same conditions, three reference models were established: BPNN, support vector regression (SVR), and GRU.

An evaluation was conducted using indicators of mean absolute error (MAE), root mean square error (RMSE), and symmetric mean absolute percent error (SMAPE) to evaluate the degree of change and accuracy of the data and to measure the predictive effect of the model. These three indicators represented the deviation of the model prediction for the real value fitting, and the smaller the result was, the more accurate it was. The calculation formula is shown in Equations (10)–(12):

$$\text{MAE}_{(y',y)} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i' - y_i \right| \tag{10}$$

$$\text{RMSE}_{(y',y)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i' - y_i \right)^2} \tag{11}$$

$$\text{SMAPE}_{(y',y)} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{\left| y_i' - y_i \right|}{\left(y_i' + y_i \right)/2} \tag{12}$$

where n is the total number of samples, y_i is the monitored value, and y_i' is the predicted value.

4.2. Network Parameters

The training of CT-VGG-GRU adopted small-batch training; the batch size was 50 and the number of iterations was 100. The dropout probability between each layer

was set to 0.1 to avoid overfitting. Then, the CT-VGG-GRU model was evaluated on the test set and calculated MAE, RMSE, and SMAPE. The hyperparameters were set for optimal predictive performance, including the time step and the number of neurons. The time step represents the size of the sliding window (i.e., how many time steps of data are used to predict the next data point). The number of neurons refers to how many neuron nodes need to be set to achieve the best prediction effect. As shown in Table 4, the number of neurons were selected from the candidate set {16, 32, 64, 128, 256}. As the number of neurons increased, the predictive performance first increased and then decreased. When the number of neurons was 64, RMSE and MAE were the lowest. Therefore, this value was used in the experiment. The number of neurons remained constant, and the time step changed. The time step was selected from the candidate set {1, 2, 3, 4, 5}. The experimental results showed that the prediction effect of the best model had a time step of 3. The candlestick chart contained three days of data and the convolution step of the candlestick chart generator was 1.

Table 4. Optimal parameters of the CT-VGG-GRU model. The best hyperparameters and experimental results are highlighted in bold.

Parameter	Set of Feasible Values	Optimal Value	MAE	RMSE
Neuron number	{16, 32, 64, 128, 256}	16	0.724	1.124
		32	0.512	0.874
		64	**0.347**	**0.547**
		128	0.478	0.812
		256	0.724	1.451
Time step	{1, 2, 3, 4, 5}	1	0.674	0.912
		2	0.475	0.624
		3	**0.241**	**0.425**
		4	0.382	0.824
		5	0.531	1.025

4.3. Prediction Performance

The indicators of water quality—DO, COD_{Mn}, and NH_3-N—were selected as the performance evaluation indicators of the established CT-VGG-GRU model in this part. The online monitoring data from January 2012 to December 2019 were used to train the model, and the daily mean values of DO, COD_{Mn}, and NH_3-N concentration from January 2020 to March 2020 were predicted. Among them, the hourly values of DO, COD_{Mn}, and NH_3-N concentration from January 2012 to December 2019 at the Yangshuo station were used as the benchmark training dataset of the VGG. After training and convergence, the optimal model weight of CT-VGG-GRU was obtained. To verify the predictive accuracy of the optimal model, the test dataset was used for model evaluation. As shown in Figure 7, the predicted values of DO, COD_{Mn}, and NH_3-N from January 2020 to March 2020 were compared with the real values. Compared with the VGG-GRU model, the proposed model showed accurate prediction performance in the whole prediction range, even when the pollutant concentration was mutated, indicating that the model could well process nonlinear characteristics and mutation of time series.

The traditional artificial intelligence model was only simple numerical fitting, and there was no reasonable explanation for the "black box" network model. However, the water pollution process was reasonably classified according to the candlestick theory, which made the prediction model more interpretable and accurate. The prediction performances of VGG-GRU and CT-VGG-GRU are shown in Figure 7. When the process information of water pollution was not effectively classified, VGG-GRU had a greater error. According to the classification law of water quality diffusion, the predicted trend of CT-VGG-GRU was more consistent with the true trend. On the basis of the candlestick theory, the water pollution process information was effectively classified, the water pollution change process was reflected, and the water pollution process information was also retained. VGG-

GRU, however, only randomly classified water pollution processes without a reasonable scientific explanation. Therefore, CT-VGG-GRU was more adaptable and better explained the information of the water pollution process, which made the predicted trend more consistent with the actual trend.
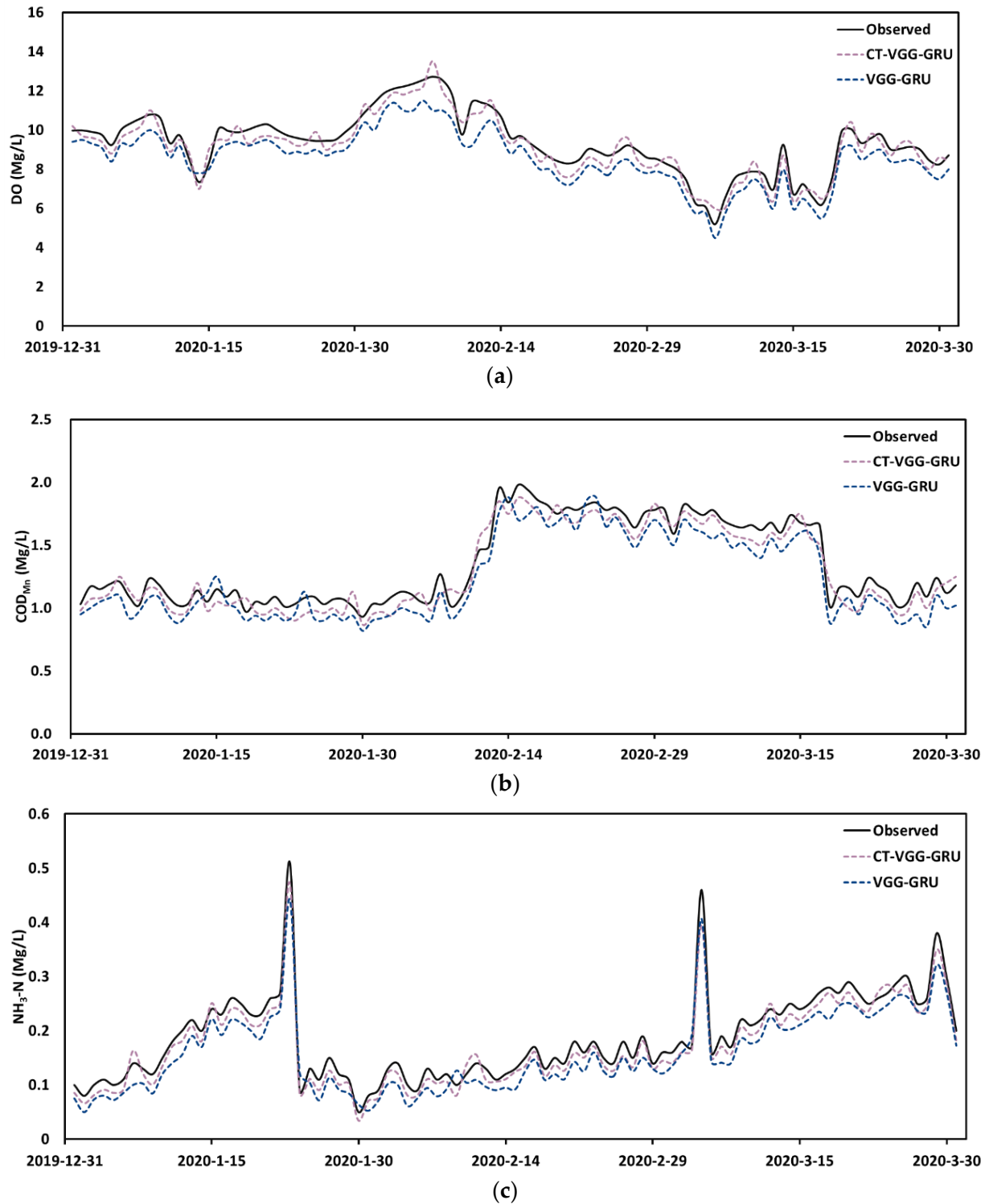

Figure 7. The predicted results of water quality indicators of the CT-VGG-GRU model and the VGG-GRU model were compared with the observed values, (**a**) DO, (**b**) COD_{Mn}, (**c**) NH_3-N.

4.4. Comparison of the Proposed Model with Other Methods

The quantitative results of MAE, RMSE, and SMAPE are listed in Table 5. Under the same conditions, BPNN, SVR, GRU, VGG-GRU, and CT-VGG-GRU were compared and analyzed. As shown in Table 5, CT-VGG-GRU obtained the best results compared with the other models. The prediction performance of BPNN and SVR was worse, and the errors were larger than other methods because the nonlinear relationship between continuous time series data on water quality could not be captured. With the same parameters, the MAE, RMSE, and SMAPE of GRU, VGG-GRU, and CT-VGG-GRU were significantly better than BPNN and SVR, which indicated that the long-term historical process of water pollution was fully reflected and the nonlinear relationship of continuous time series was better captured. The CT-VGG-GRU model that coupled the candlestick chart and deep learning had higher water quality prediction accuracy than GRU and VGG-GRU. Compared with all baselines, CT-VGG-GRU had the best performance because the future trend of water quality indicators was well predicted. At the same time, the results reflected a good effect on the prediction of extreme values, and better adaptability.

Table 5. Comparison of model performance. The best experimental results of the CT-VGG-GRU model are highlighted in bold.

Indicator	Method	MAE	RMSE	SMAPE (%)
DO	BPNN	1.121	1.195	0.093
	SVR	0.810	0.902	0.066
	GRU	0.464	0.520	0.037
	VGG-GRU	0.324	0.375	0.029
	CT-VGG-GRU	**0.284**	**0.315**	**0.022**
COD_{Mn}	BPNN	0.494	0.511	0.586
	SVR	0.347	0.364	0.380
	GRU	0.203	0.219	0.209
	VGG-GRU	0.141	0.157	0.137
	CT-VGG-GRU	**0.113**	**0.122**	**0.108**
NH_3-N	BPNN	0.057	0.061	0.606
	SVR	0.041	0.046	0.417
	GRU	0.027	0.029	0.241
	VGG-GRU	0.018	0.021	0.178
	CT-VGG-GRU	**0.014**	**0.016**	**0.127**

Under the same conditions, the DO, COD_{Mn}, and NH_3-N concentrations predicted by different models including BPNN, SVR, GRU, and CT-VGG-GRU are shown in Figure 8. The online monitoring data from January 2012 to March 2020 were used to train the model, and the daily mean values of DO, COD_{Mn}, and NH_3-N concentration from April 2020 to June 2020 were predicted. Among them, the hourly value of DO, COD_{Mn}, and NH_3-N concentration from January 2012 to March 2020 at the Yangshuo station was used as the benchmark training dataset of VGG. Through comparative analysis, the proposed method was superior to the mainstream methods such as SVR and GRU. The higher prediction accuracy than that of other methods might be due to two reasons. First, the strong robustness, memory ability, self-learning ability, and nonlinear mapping ability of the neural network were helpful to better predict water quality indicators. Second, the deep learning model based on the candlestick theory was sensitive to local change information, which improves the accuracy of extreme value prediction.

To show the prediction performances of DO, COD_{Mn}, and NH_3-N intuitively, the observed and predicted scatter plots of different models including BPNN, SVR, GRU, and CT-VGG-GRU are shown in Figure 9. The predicted results of the proposed model were consistent with the trend of the observed values. The scatter plot identifies the trend that the model underestimates high concentration and overestimates low concentration [40]. The observed values of BPNN and the shallow machine learning model (SVR) could be seen to differ more from the predicted values than the deep learning models, which

indicated that they failed to consider the impact of time series water quality data on prediction results (the impact of the previous time on the next time). In contrast, the GRU and CT-VGG-GRU models obviously achieved better prediction performance in water quality prediction. Therefore, that time series neural network models had higher prediction accuracy in the application of water pollution prediction. Compared with all of the noted models, the CT-VGG-GRU model was closer to the real trend of water quality change and more sensitive to local sharp changes in water quality pollutants, mainly because the VGG model based on the candlestick chart was able to obtain richer change information. The established CT-VGG-GRU model in the paper could provide a new method for water quality prediction.
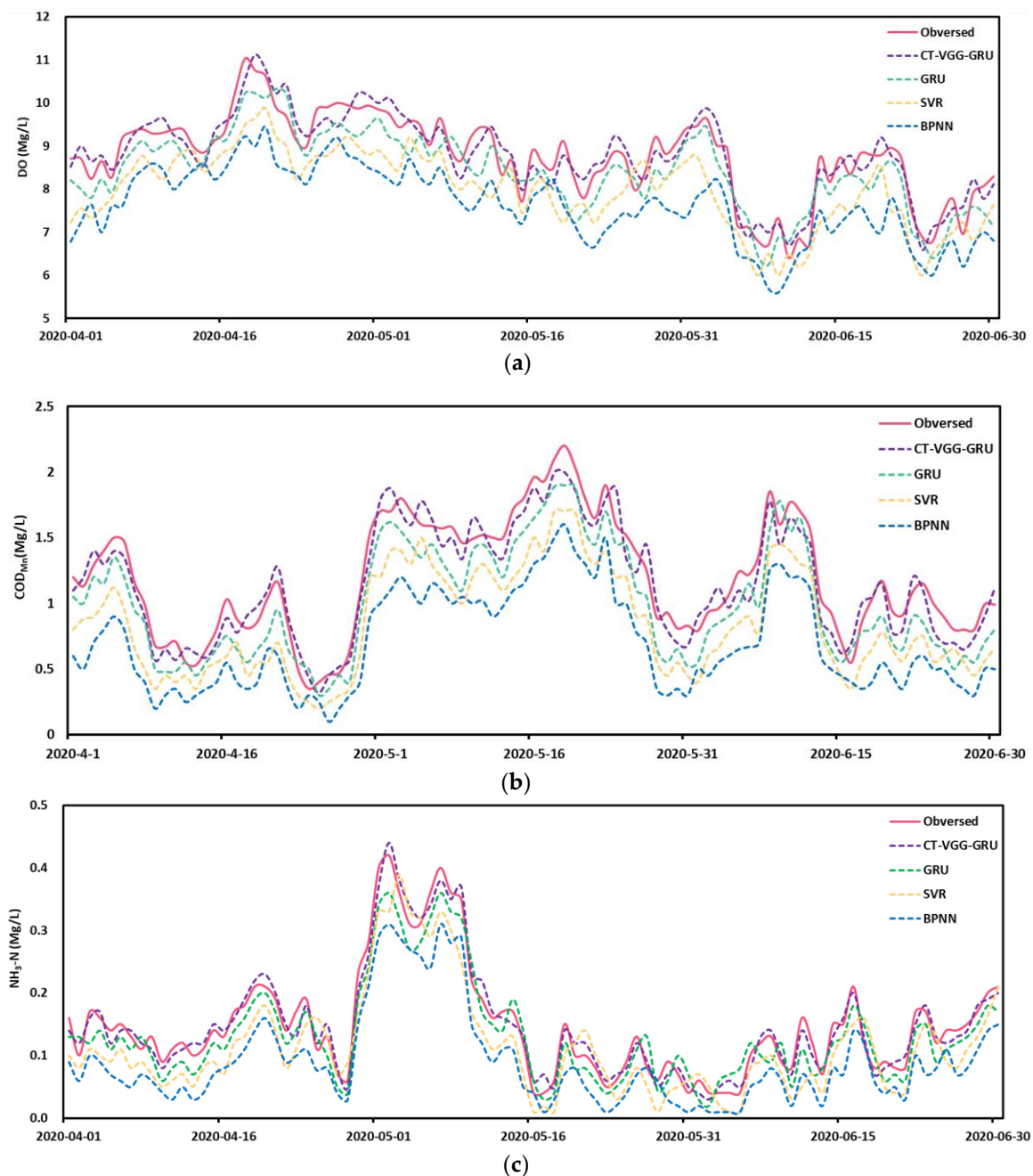
(a)

(b)

(c)

Figure 8. The predicted results of water quality indicators each model were compared with the observed values, (**a**) DO, (**b**) COD_{Mn}, (**c**) NH_3-N.
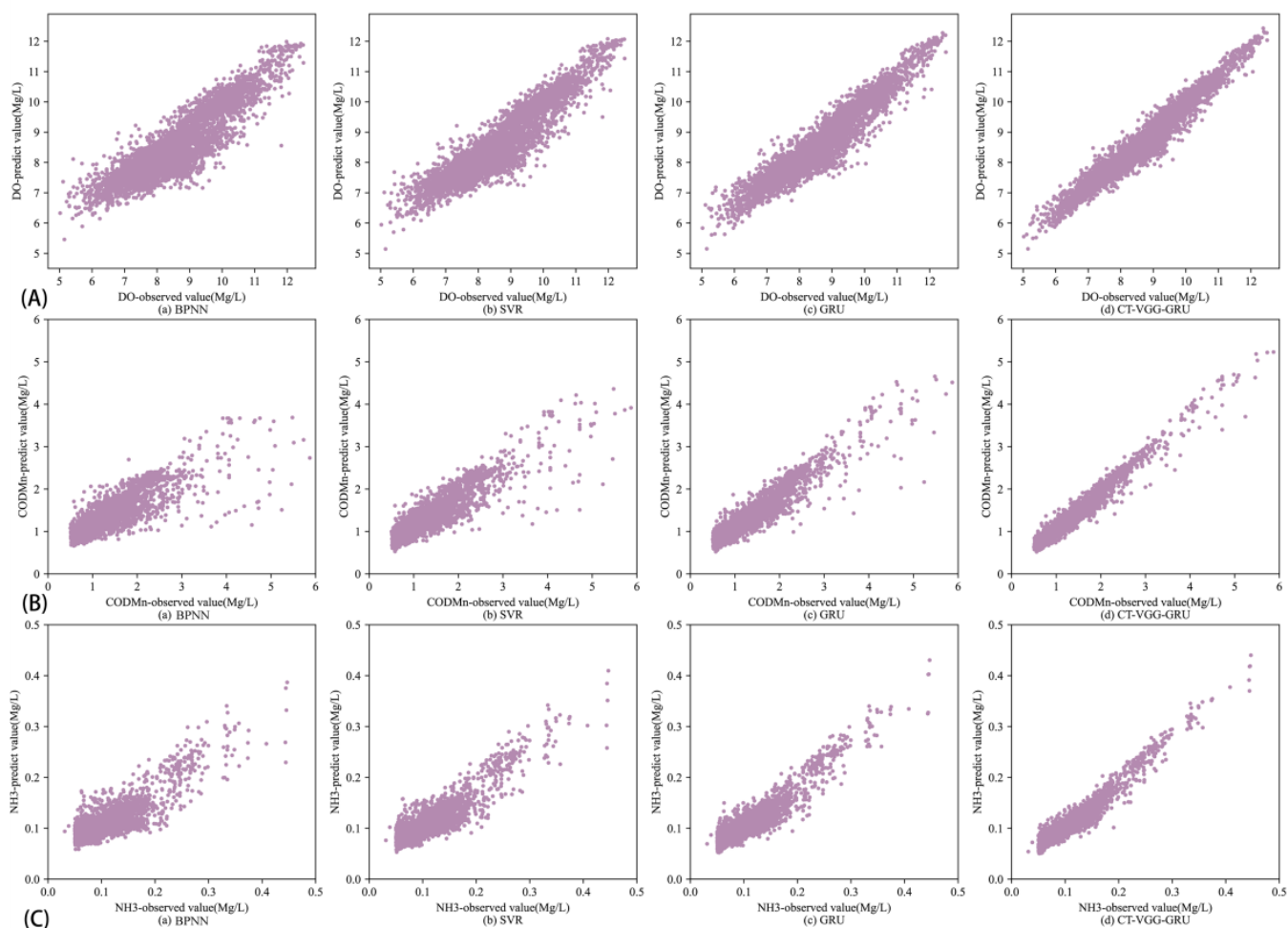
Figure 9. Scatter plots of comparison models: (**A**) DO, (**B**) COD$_{Mn}$, (**C**) NH$_3$-N.

5. Discussion

Improving the prediction accuracy of water quality parameters is an important and arduous task for water pollution management decision makers. At present, there are many water quality prediction models, but understanding how to improve the effectiveness and accuracy of the prediction models remains difficult. In the prediction and simulation methods of water pollution, common traditional physical models had problems, such as difficulty in parameter calibration, lack of basic data support, and complex model construction. On the basis of massive time series data, machine learning and deep learning models simply fitted the numerical curve with the network structure, which was not fully connected with the transmission process of water pollution. Therefore, the interpretability was not sufficient, and the prediction accuracy was not satisfactory. In the water environment, the water pollution process was affected by various factors, such as complex transport process and climate change. The changing trend of pollutant concentration was obvious, and the classification types of different pollution processes were presented. Therefore, a water pollution prediction model based on candlestick theory and deep learning was proposed. The experimental training data were constructed from a group of time series data, including pollutant data and hydrometeorological data. On this dataset, the CT-VGG-GRU model was trained to describe the classification characteristics of the water pollution process. First, based on the candlestick theory, the water pollution process was classified by graphics. Then, the classification characteristics of the water pollution process were extracted according to VGG. Finally, combined with

the other scenario parameters of each station, the water pollution was predicted by the time series deep learning model, GRU.

In the experimental validation part, the performance of the model was evaluated and compared with the BPNN, SVR, and GRU models. Water quality indicators—DO, COD_{Mn}, and NH_3-N—were selected as the performance evaluation indicators. The mean absolute errors (MAE) were 0.284, 0.113, and 0.014; the root mean square errors (RMSE) were 0.315, 0.122, and 0.016; and the symmetric mean absolute percentage errors (SMAPE) were 0.022, 0.108, and 0.127, respectively, which were superior to the machine learning model and the simple deep learning model. These results indicated that the concentration of water pollutants could be effectively predicted by the proposed model. Compared with the VGG-GRU model, the changing law of continuous time series data could be reflected, and the periodic characteristics and time series characteristics of water quality data could be fully extracted. Based on the graphical form of the candlestick chart, the water pollution process was effectively classified. From the physical equation of diffusion, the pollution process in different periods was reflected, and the different trends of water pollution could be effectively classified. Therefore, the proposed model based on the candlestick chart was more explanatory for the water pollution process, and the prediction accuracy was higher. Additionally, the proposed model could not only predict the daily mean value of water pollutant concentration, but also the daily maximum and minimum values, etc., and could be used to predict the atmospheric pollutant concentration.

According to these results, this study provided a new method for the prediction of water pollution and a strong basis for water quality prediction, thereby helping to save time and labor and improve human health. However, in the study of the CT-VGG-GRU hybrid model, a gap exists between the predicted value and the actual value of pollutant concentration. Because water pollution transmission is an extremely complex environmental science phenomenon, it is affected by various factors such as the chemical process and the geographical environment. If the chemical change process and geographical factors can be analyzed during the diffusion of water pollution, the errors will be limited to within a small range, and the prediction results will be closer to the real value. At present, there are actually many deep learning models for similar problems in the environmental domain. While considering various influencing factors such as chemical process and geographical environment, other deep learning models not considered in this paper can also be tried, and there may be a model that is more suitable for this research. All these are future research directions.

Author Contributions: Conceptualization, R.X.; methodology, R.X. and H.W.; software, R.X., H.W. and W.W.; validation, Y.C. and Q.Z.; formal analysis, Y.C. and Q.Z.; investigation, J.L. and S.S.; resources, J.L. and S.S.; data curation, W.W.; writing—original draft preparation, W.W.; writing—review and editing, R.X., H.W. and W.W.; visualization, J.L.; supervision, S.S. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no competing interest.

References

1. Li, F.; Zhang, G.; Xu, Y.J. Separating the impacts of climate variation and human activities on runoff in the Songhua River Basin, Northeast China. *Water* **2014**, *6*, 3320–3338. [CrossRef]
2. Wan, H.; Mao, Y.; Cai, Y.; Li, R.; Feng, J.; Yang, H. An SPH-based mass transfer model for simulating hydraulic characteristics and mass transfer process of dammed rivers. *Eng. Comput.* **2021**, *38*, 3169–3184. [CrossRef]

3. Akoko, G.; Le, T.H.; Gomi, T.; Kato, T. A review of SWAT model application in Africa. *Water* **2021**, *13*, 1313. [CrossRef]
4. Aawar, T.; Khare, D. Assessment of climate change impacts on streamflow through hydrological model using SWAT model: A case study of Afghanistan. *Model. Earth Syst. Environ.* **2020**, *6*, 1427–1437. [CrossRef]
5. Ramteke, G.; Singh, R.; Chatterjee, C. Assessing impacts of conservation measures on watershed hydrology using MIKE SHE model in the face of climate change. *Water Resour. Manag.* **2020**, *34*, 4233–4252. [CrossRef]
6. Mbuh, M.J.; Mbih, R.; Wendi, C. Water quality modeling and sensitivity analysis using Water Quality Analysis Simulation Program (WASP) in the Shenandoah River watershed. *Phys. Geogr.* **2019**, *40*, 127–148. [CrossRef]
7. Kouadri, S.; Kateb, S.; Zegait, R. Spatial and temporal model for WQI prediction based on back-propagation neural network, application on EL MERK region (Algerian southeast). *J. Saudi Soc. Agric. Sci.* **2021**, *20*, 324–336. [CrossRef]
8. Wang, X.; Wang, K.; Ding, J.; Chen, X.; Li, Y.; Zhang, W.J. Predicting water quality during urbanization based on a causality-based input variable selection method modified back-propagation neural network. *Environ. Sci. Pollut. Res.* **2021**, *28*, 960–973. [CrossRef]
9. Liang, N.; Zou, Z.; Wei, Y. Regression models (SVR, EMD and FastICA) in forecasting water quality of the Haihe River of China. *Desalination Water Treat.* **2019**, *154*, 147–159. [CrossRef]
10. Su, X.; He, X.; Zhang, G.; Chen, Y.; Li, K. Research on SVR Water Quality Prediction Model Based on Improved Sparrow Search Algorithm. *Comput. Intell. Neurosci.* **2022**, *2022*, 7327072. [CrossRef]
11. Wang, Y.; Yuan, Y.; Pan, Y.; Fan, Z. Modeling daily and monthly water quality indicators in a canal using a hybrid wavelet-based support vector regression structure. *Water* **2020**, *12*, 1476. [CrossRef]
12. Hassanjabbar, A.; Nezaratian, H.; Wu, P. Climate change impacts on the flow regime and water quality indicators using an artificial neural network (ANN): A case study in Saskatchewan, Canada. *J. Water Clim. Chang.* **2022**, *13*, 3046–3060. [CrossRef]
13. Prasad, D.V.V.; Venkataramana, L.Y.; Kumar, P.S.; Prasannamedha, G.; Harshana, S.; Srividya, S.J.; Harrinei, K.; Indraganti, S. Analysis and prediction of water quality using deep learning and auto deep learning techniques. *Sci. Total Environ.* **2022**, *821*, 153311. [CrossRef]
14. Wan, H.; Xu, R.; Zhang, M.; Cai, Y.; Li, J.; Shen, X. A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods. *J. Hydrol.* **2022**, *612*, 128081. [CrossRef]
15. Li, L.; Jiang, P.; Xu, H.; Lin, G.; Guo, D.; Wu, H. Water quality prediction based on recurrent neural network and improved evidence theory: A case study of Qiantang River, China. *Environ. Sci. Pollut. Res.* **2019**, *26*, 19879–19896. [CrossRef]
16. Liu, Y.; Zhang, Q.; Song, L.; Chen, Y. Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. *Comput. Electron. Agric.* **2019**, *165*, 104964. [CrossRef]
17. Liang, Z.; Zou, R.; Chen, X.; Ren, T.; Su, H.; Liu, Y. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. *J. Hydrol.* **2020**, *581*, 124432. [CrossRef]
18. Barzegar, R.; Aalami, M.T.; Adamowski, J. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 415–433. [CrossRef]
19. Wan, H.; Tan, Q.; Li, R.; Cai, Y.; Shen, X.; Yang, Z.; Shen, X. Incorporating Fish Tolerance to Supersaturated Total Dissolved Gas for Generating Flood Pulse Discharge Patterns Based on a Simulation-Optimization Approach. *Water Resour. Res.* **2021**, *57*, e2021WR030167. [CrossRef]
20. Xu, R.; Deng, X.; Wan, H.; Cai, Y.; Pan, X. A deep learning method to repair atmospheric environmental quality data based on Gaussian diffusion. *J. Clean. Prod.* **2021**, *308*, 127446. [CrossRef]
21. Xie, H.; Zhao, X.; Wang, S. A comprehensive look at the predictive information in Japanese candlestick. *Procedia Comput. Sci.* **2012**, *9*, 1219–1227. [CrossRef]
22. Cagliero, L.; Fior, J.; Garza, P. Shortlisting machine learning-based stock trading recommendations using candlestick pattern recognition. *Expert Syst. Appl.* **2023**, *216*, 119493. [CrossRef]
23. Lan, Q.; Zhang, D.; Xiong, L. Reversal pattern discovery in financial time series based on fuzzy candlestick lines. *Syst. Eng. Procedia* **2011**, *2*, 182–190. [CrossRef]
24. Tsai, C.-F.; Quan, Z.-Y. Stock prediction by searching for similarities in candlestick charts. *ACM Trans. Manag. Inf. Syst.* **2014**, *5*, 1–21. [CrossRef]
25. Lee, K.; Jo, G. Expert system for predicting stock market timing using a candlestick chart. *Expert Syst. Appl.* **1999**, *16*, 357–364. [CrossRef]
26. Chen, S.; Bao, S.; Zhou, Y. The predictive power of Japanese candlestick charting in Chinese stock market. *Phys. A Stat. Mech. Its Appl.* **2016**, *457*, 148–165. [CrossRef]
27. Hung, C.-C.; Chen, Y.-J. DPP: Deep predictor for price movement from candlestick charts. *PLoS ONE* **2021**, *16*, e0252404. [CrossRef]
28. Xu, R.; Liu, X.; Wan, H.; Pan, X.; Li, J. A Feature Extraction and Classification Method to Forecast the PM2. 5 Variation Trend Using Candlestick and Visual Geometry Group Model. *Atmosphere* **2021**, *12*, 570. [CrossRef]
29. Liang, M.; Wu, S.; Wang, X.; Chen, Q. A stock time series forecasting approach incorporating candlestick patterns and sequence similarity. *Expert Syst. Appl.* **2022**, *205*, 117595. [CrossRef]
30. Nison, S. *Japanese Candlestick Charting Techniques*; New York Institute of Finance: New York, NY, USA, 1991.
31. Hu, W.; Si, Y.-W.; Fong, S.; Lau, R.Y.K. A formal approach to candlestick pattern classification in financial time series. *Appl. Soft Comput.* **2019**, *84*, 105700. [CrossRef]
32. Farahbod, F. Mathematical investigation of diffusion and decomposition of pollutants as a basic issue in water stream pollution. *Arab. J. Geosci.* **2020**, *13*, 918. [CrossRef]

33. Zhuang, W.; Pang, Y.; Lv, J. Research on the Integration of Two-dimensional Water Quality Model and Geographic Information System. In Proceedings of the 2007 Major Water Conservancy and Hydropower Science and Technology Frontier Academician Forum and the First China Water Conservancy Doctoral Forum Proceedings, Nanjing, China, 11 November 2007; pp. 557–563.

34. Wang, H.; Yan, W.; Wang, J.; Duan, W. Exploring Distribution Rules and Variation Trends of Precipitation in the Upper Lijiang River from 1951 to 2016, Guangxi Province, China. *J. Coast. Res.* **2020**, *105*, 1–5. [CrossRef]

35. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]

36. Yang, Y.; Xiong, Q.; Wu, C.; Zou, Q.; Yu, Y.; Yi, H.; Gao, M. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. *Environ. Sci. Pollut. Res.* **2021**, *28*, 55129–55139. [CrossRef]

37. Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1309–1322. [CrossRef]

38. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]

39. Wang, Y.; Liao, W.; Chang, Y. Gated recurrent unit network-based short-term photovoltaic forecasting. *Energies* **2018**, *11*, 2163. [CrossRef]

40. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **2020**, *721*, 137612. [CrossRef]