AQUA

Water Infrastructure, Ecosystems and Society

© 2022 The Authors AQUA — Water Infrastructure, Ecosystems and Society Vol 71 No 9, 963 doi: 10.2166/aqua.2022.036

A combined water quality pollution prediction model based on the Spark big data platform

Zhihui Sun and Yiqing Fan*

School of Economics and Trade, Fujian Jiangxia University, Fuzhou, Fujian, China

*Corresponding author. E-mail: louyao0340@163.com

ABSTRACT

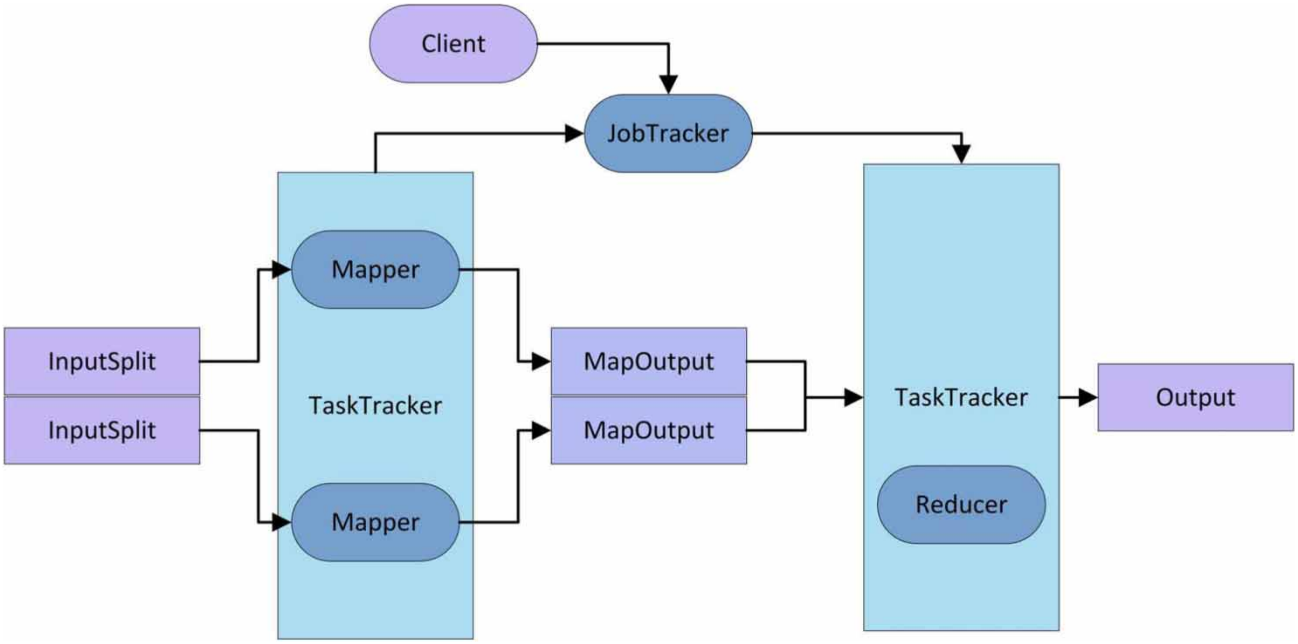
Water quality prediction is the basic work of water resource management and pollution control, and it is crucial to accurately predict the trend of pollutant concentration in water bodies over time. Water quality data prediction has an important significance, as it provides data support for the effective estimation of water quality, and is also an indirect way to protect water resources and the environment. At present there are a variety of water quality prediction methods, but these methods still have some shortcomings. In this paper, the main water quality pollution indicators such as the dissolved oxygen (DO), ammonia nitrogen (NH_3-N) and total phosphorus (P) data were the object of study to build a water quality prediction model. The water quality prediction index contains numerous nonlinear correlation characteristics that results in low training efficiency on a large-scale data. Therefore, a combined water quality prediction model based on integrated ensemble empirical mode decomposition (EEMD) and cascade support vector machine (Cascade SVM) is proposed. First, the EEMD method is used to highlight the real characteristics of the original water quality data series. Then, the parallel training and prediction process are realized by the Spark, a distributed computing engine, to parallelize the traditional Cascade SVM. The experimental results show that the proposed combined model shows a strong superiority in many aspects of performance such as training efficiency and prediction accuracy.

Key words: empirical mode decomposition, parallel computing, predictive model, support vector machine, water pollution

HIGHLIGHTS

- Proposes a combined water quality prediction model based on EEMD and Cascade SVM.
- Improves the accuracy of the prediction results.
- A combined water quality prediction model proposed in this paper has a higher accuracy.
- A combined water quality prediction model proposed in this paper has less prediction time.
- The proposed combined model shows a strong superiority in training efficiency and prediction accuracy.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Due to the unreasonable living habits and production methods of human beings, the problem of water pollution has become increasingly serious worldwide, exceeding the maximum pollution load that the natural environment can bear. Water quality degradation is considered one of the most serious environmental problems worldwide, as it can destroy the ecological balance of water bodies and endanger regional environmental security. Therefore, how to accurately predict water quality is of great importance for both social and economic development. In recent years, with the high frequency of water pollution events, water quality prediction has gradually become a popular issue of concern for environmental management departments in many countries (regions) (Qaderi & Engineering 2017; Ding *et al.* 2019; Rashid *et al.* 2021; Singha *et al.* 2021).

Water quality forecasting is the basic work of water resources and environmental management, the use of scientific and reasonable water quality forecasting methods, in order to accurately reflect the current water quality and pollution, to clarify the development of water quality changes in the law, and to identify the main pollution problems. The purpose of water quality prediction is to prevent the further deterioration of water quality. With the help of monitoring the data obtained and the relevant information collected, people in the current water quality depend on the future development trend of water quality to make scientific predictions and inferences, in order to propose corresponding improvement methods. At this stage, the main methods of water quality prediction can be divided into two types (Zhang *et al.* 2016; Avila *et al.* 2017; Khadr & Elshemy 2017): (1) prediction methods based on classical statistical analysis and (2) prediction methods based on artificial intelligence modeling. Shi & Zou (2014) used probability distributions of independent residuals to generate synthetic water quality data and used autoregressive integrated moving average (ARIMA) models to predict future water quality data for complex waters. Park & Koo (2015) proposed an ARIMA model-based water quality prediction model that can predict common monitoring indicators such as dissolved oxygen (DO) and ammonia nitrogen (NH_3-N). Katimon *et al.* (2018) used the ARIMA model to model the water quality data of the Johor River to achieve accurate prediction of hydrological variables. Zhai *et al.* (2021) used statistical analysis methods to achieve the prediction of hazardous chemical accidents in drinking water sources in the Three Gorges reservoir area.

Since most statistical-based water quality prediction methods are normally distributed, they cannot be applied to other river waters. With continuous research, artificial intelligence methods have been rapidly developed in the field of water quality prediction. Setshedi *et al.* (2021) used artificial neural networks (ANNs) to predict the red tide phenomenon in a water quality dataset. Hrnjica and Bonacci (2019) use feedforward and recurrent neural networks for the lake water level prediction. In addition to ANNs, many researchers have attempted to use various machine learning techniques for multiple water quality metrics prediction (Shihab & Al-Tayyar 2019; Wang *et al.* 2019; Yi *et al.* 2019; Deng *et al.* 2021; Searcy & Boehm 2021). For example, Cao *et al.* (2020) proposed a genetic algorithm-optimized support vector machine (SVM) to predict future water quality conditions. However, optimal support vector machines are less efficient to train on large-scale water quality data.

Therefore, in order to solve the above problems, a combined water quality prediction model based on ensemble empirical mode decomposition (EEMD) and cascade support vector machine (Cascade SVM) is proposed in this paper. First, DO, NH_3-N, and total phosphorus (P) from water quality monitoring data are selected as predictors (Haleem *et al.* 2019; Hu *et al.* 2019; Harun *et al.* 2020), and EEMD is used to decompose water quality time series data to obtain relatively realistic components. Second, Cascade SVM, as a parallel SVM, can improve its own training efficiency on a large-scale data through global problem decomposition, filtering and feedback. Therefore, in this paper, we try to use Spark-based parallelized Cascade SVM for each component obtained from the decomposition to make predictions. Finally, the corresponding outputs from the above process are combined to obtain the water quality prediction results of the combined model.

The rest of the paper is organized as follows: In Section 2, the EEMD method for water quality data series is studied in detail, while Section 3 provides the detailed data of the proposed combined water quality prediction model. Section 4 provides the results and discussion. Finally, the paper is concluded in Section 5.

2. EEMD METHOD FOR WATER QUALITY DATA SERIES

2.1. Principle of empirical mode decomposition method

Empirical mode decomposition (EMD) is usually used to deal with non-smooth nonlinear signal sequences (Chen *et al.* 2018; Du *et al.* 2018; Lu *et al.* 2018). Generally speaking, most of the intrinsic decomposition methods are only suitable for data with certain fixed characteristics. For example, wavelet transform decomposition methods require the decomposed data to be smooth and linear. On the contrary, Fourier transform decomposition is mainly used to deal with smooth cyclic data cases. EMD takes part of the global signal as the basis to resolve the signal tendency or fluctuation pattern, and also generates several intrinsic mode functions (IMFs). Theoretically speaking, EMD can be used for different types of signal analysis.

EMD resolves the original data series $s(t)$ into a combination of several IMFs and a residual of the form

$$s(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{1}$$

where n represents the number of IMFs, $c_i(t)$ represents the ith IMF and $r_n(t)$ represents the nth residual.

2.2. Process of EEMD

The homogeneous distribution of the Gaussian white noise spectrum is utilized to drive the signals of different time scales to actively disperse to a suitable reference standard. In the initial signal to add the Gaussian white noise, on the one hand, can provide a uniform distribution of the signal resolution standard; on the other hand, it can smooth the interference of the pulse, and thus is superior to highlight the real characteristics of the initial water quality sequence. The origin of EEMD is a repeated addition of the Gaussian white noise to the multiple EMD, and the detailed steps of the process are as follows:

- Add a Gaussian white noise sequence $n_m(t)$ with 0 mean to the original signal to be analyzed $s(t)$.

$$s_m(t) = s(t) + n_m(t) \tag{2}$$

- Resolving the signal after adding the Gaussian white noise sequence into a set of IMFs by EMD.
- Adding a different Gaussian white noise sequence each time, and then repeating the above steps several times.
- Calculate the mean value of the decomposed IMF by using the principle of uncorrelated random sequences, so as to suppress the influence of the Gaussian white noise on the real IMF. The final IMF that can be obtained by EEMD parsing is:

$$c_i(t) = \frac{\sum_{m=1}^{N} c_{i,m}(t)}{N} \tag{3}$$

where N is the number of EMD integrations and $c_{i,m}(t)$ is the ith IMF obtained from the mth EMD.

- Assuming an infinite value of N in Step 4. The result of the EEMD resolution can be expressed as:

$$s(t) = \sum_{i=1}^{n} c_i(t) + \bar{r} \tag{4}$$

where $\bar{r}$ is the mean of the residuals.

The decomposition flow of EEMD is shown in Figure 1.

3. PROPOSED COMBINED WATER QUALITY PREDICTION MODEL

3.1. Basic framework of the model

In many water quality prediction literatures, single model prediction is widely used, such as SVM, gray model prediction method and ANN, but the applications of combined model prediction are relatively few. This paper proposes a new combined model, and through empirical research shows that the combined model in water quality prediction has a high prediction accuracy. In this paper, three indexes, DO, NH_3-N and P, in water quality data are selected as prediction indexes, and a combined water quality prediction model is constructed.

The basic framework of this model is shown in Figure 2. On the one hand, the initial water quality data time series is preprocessed by EEMD, so as to obtain a group of relatively stable components from high frequency to low frequency. On the other hand, parallel Cascade SVMs based on the Spark are used to predict each component. Finally, the prediction results of all components are integrated together to get the final water quality prediction results of the combined model.

The EEMD–Spark–Cascade SVM combined model is built by the following steps:

- Determine the optimal input structure of the model.
- Establish the training sample of the model according to the determined optimal input structure, and train and test the model by the determined training sample until the error is minimized.
- According to the constructed combined model predict the water quality data.

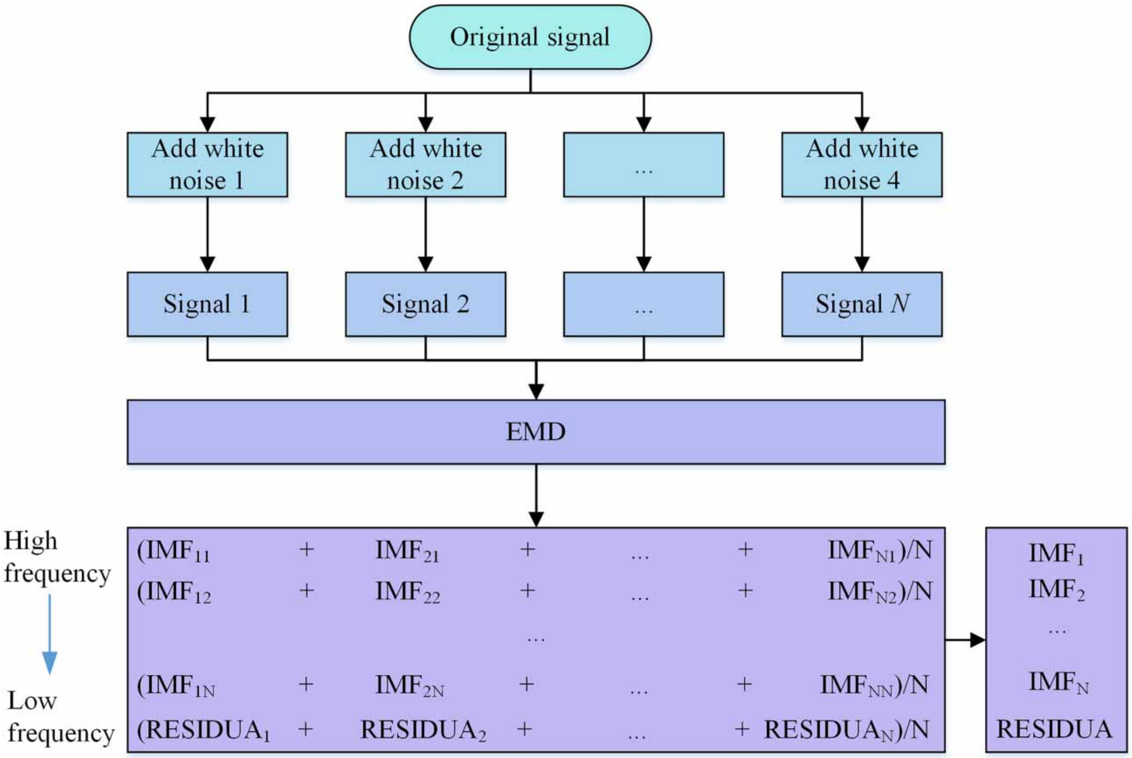


Figure 1 | Decomposition process of EEMD.

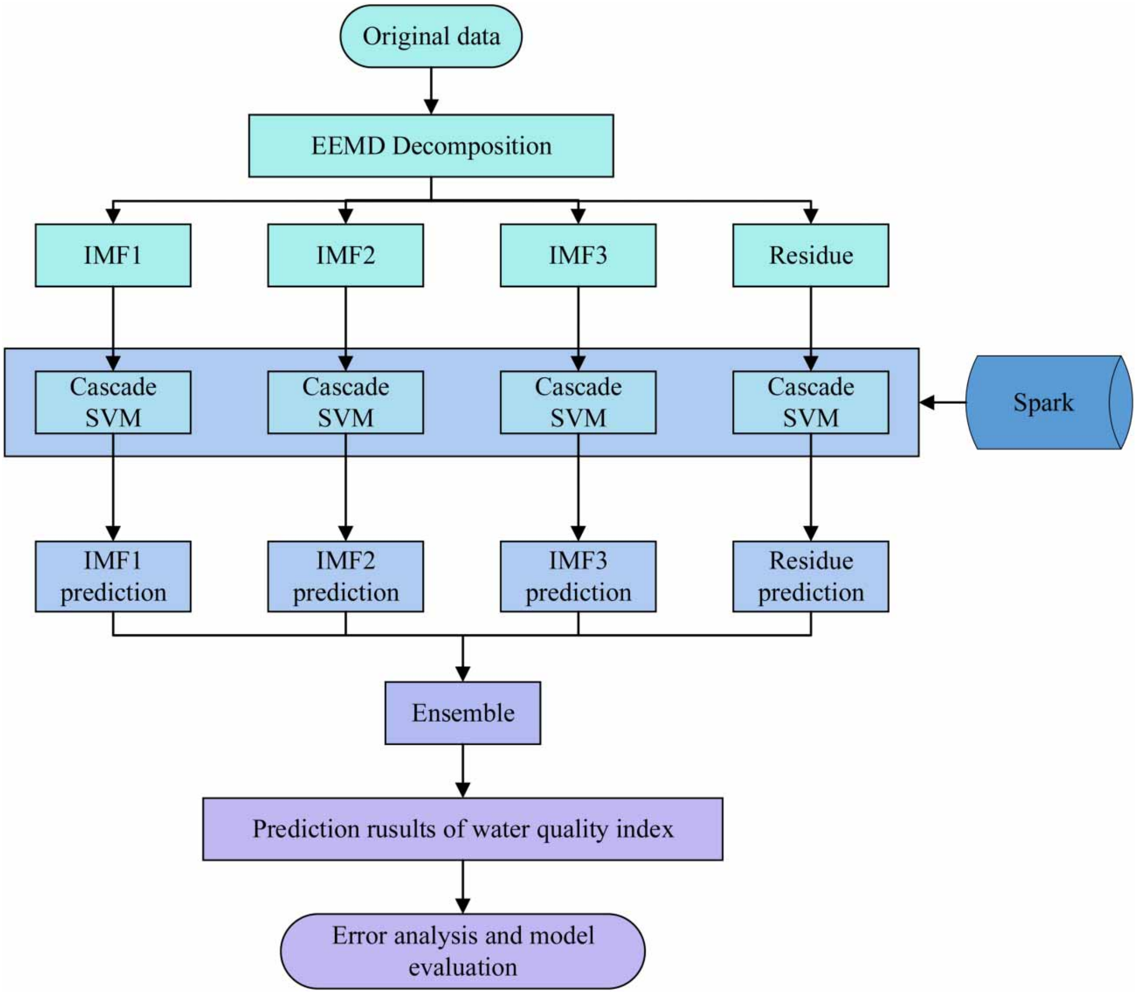


Figure 2 | Basic framework of a combined water quality prediction model.

3.2. Spark-based Cascade SVM

Cascade SVM is a parallel framework for solving SVMs designed for distributed systems (Wu & Meng 2006; Cheng & Jhan 2013; Mazo *et al.* 2017), which aims to alleviate the pressure of the solving process on memory when training large-scale datasets. Cascade SVM accelerates the training process of convergence to the global optimum. We use the Spark system for the training and prediction process of the Cascade SVM, which implements the parallelization of the traditional Cascade SVM. In order to efficiently access the water quality data under the Spark system, the working nodes of Spark are deployed on each DataNode node of Hadoop Distributed File System (HDFS). TaskTraker performs Map tasks as well as Reduce tasks. The design idea of the Map-Reduce model is shown in Figure 3.

First, the entire training set on HDFS (Huang *et al.* 2017) is divided randomly and equally, which is called Initial Random Partition (IRP). In this step, the entire training set is randomly partitioned into m ($m = 2^i$, $i = 1, 2, \ldots, N$) subsets, each corresponding to a Split block. An Receiver Register Disable (RDD) is assigned to each subset and the partition is set so that each subset corresponds to a partition.

Then, the training process is performed on the partitioned RDDs, i.e., a SVM training process is executed for each partition of the dataset. Spark starts multiple Executor processes (Zhang *et al.* 2021) on each Worker node of the cluster to complete the training process for each subset.

After training all the subsets in the previous layer in parallel, the support vectors (SVs) of each subset are merged by RDD merging operation and persisted to HDFS as input for the next layer.

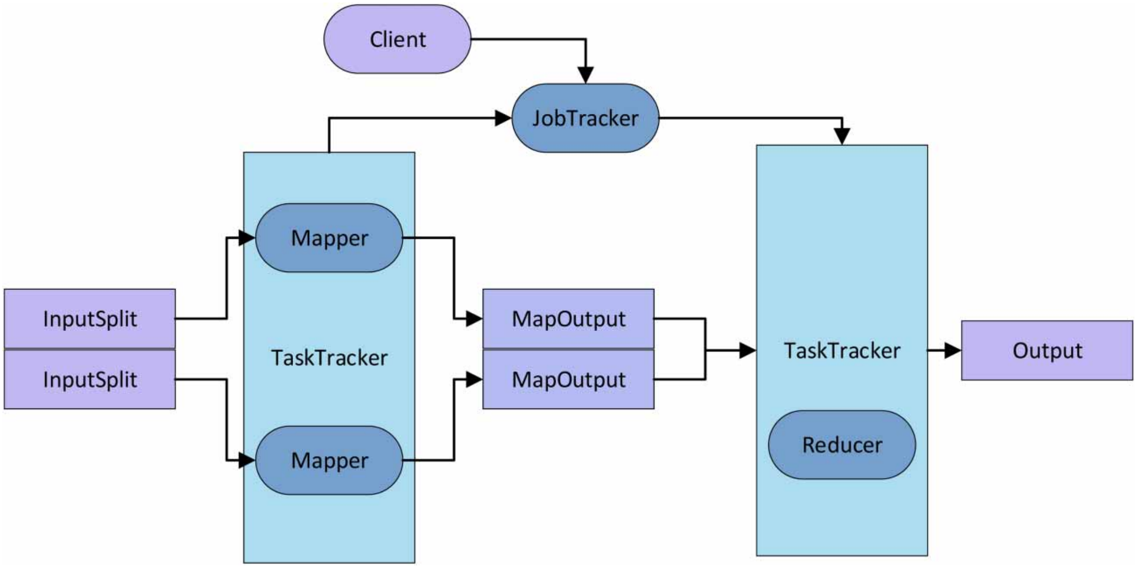


Figure 3 | Map-Reduce model design ideas.

When two SVMs are combined, TD_1 and TD_2 represent the two datasets, SVM_1 and SVM_2 represent the SVMs trained on the two datasets, $(w^*_{TD_1}, b^*_{TD_1})$ and $(w^*_{TD_2}, b^*_{TD_2})$ represent the respective classification hyperplanes and SV_1 and SV_2 represent the respective support vector. In the structure of Cascade SVM, TD_1 and TD_2 can be considered as the training and testing sets, respectively. Suppose the solution of the pairwise problem on the whole data set is $\alpha = \left(\alpha_1, \quad \alpha_2, \quad \alpha_3, \ \ldots, \ \alpha_{|TD_1|+|TD_2|}\right)$. The whole training process is to filter out the points that satisfy Equation (5).

$$y(w^*_{TD_2} \cdot x + b^*_{TD_2}) \leq 0, \quad (x, y) \in TD_1 \tag{5}$$

The prediction process for each SVM in the Cascade SVM framework is relatively simple with respect to the training process.

$$F(x) = w^*\ x + b^* = \sum_{i=1}^{N_S} \alpha_i y_i K(x_i, x) + b^* \tag{6}$$

where x denotes the point to be predicted, (w^*, b^*) denotes the SVM model on a certain training set, N_S denotes the number of SVs on that model, $K(x_i, x)$ denotes the kernel function value of the SVM, and α_i denotes the Lagrangian coefficient corresponding to the SVM.

For each point to perform the prediction process, they share an SVM and the corresponding Lagrangian solution set. The prediction process of each sample point is not affected by other sample points, which corresponds exactly to the programming model of RDD on the Spark and can be parallelized to reduce the time needed for prediction. The implementation flow of Spark-based Cascade SVM is shown in Figure 4.

4. EXPERIMENT AND RESULT ANALYSIS

4.1. Experimental environment and data sources

Spark's distributed file system HDFS was installed in the experimental environment and Spark clustering configuration was performed. All services were deployed on three Linux virtual machines in VMWare. The operating system is CentOS-6.8. The virtual machine environment configuration is shown in Table 1.

Spark services run on the Java virtual machine. The software versions used in the experiments are shown in Table 2. After all the configurations are completed, the configuration of HDFS and Spark cluster is completed by distributing the packages to all nodes.

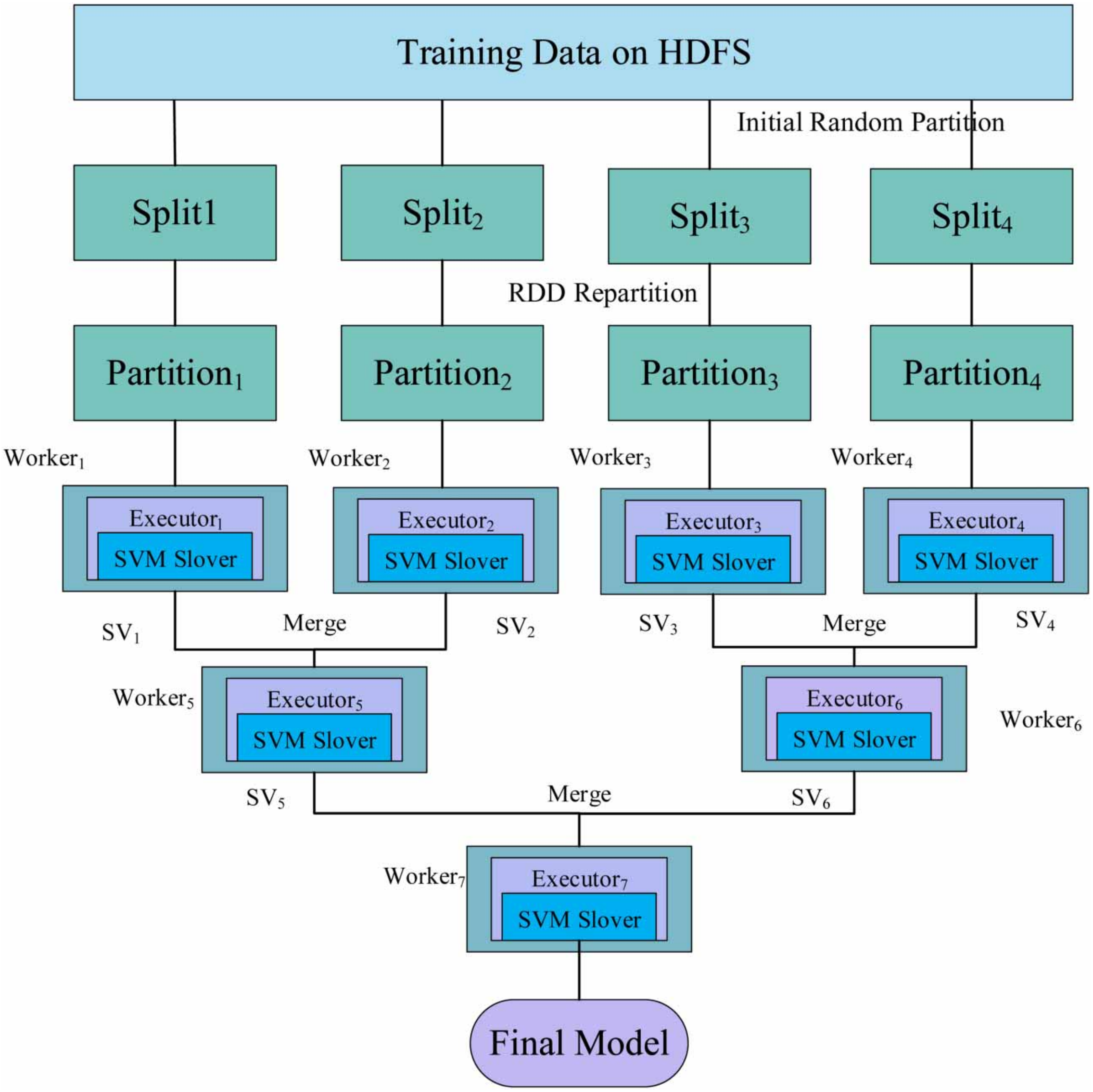


Figure 4 | Implementation flow of the Spark-based Cascade SVM.

Table 1 | Virtual machine environment configuration

No.	Hosts	Operating system	CPU core	Memory (GB)	Hard disk (GB)
1	Node1	CentOS-6.8	2	2	20
2	Node2	CentOS-6.8	4	4	100
3	Node3	CentOS-6.8	4	4	100

The experimental data were obtained from the Chuanyang River water quality site in Shanghai Taihu Lake Basin on the China Environmental Monitoring website (http://www.cnemc.cn/). The monthly data of DO, NH_3-N and P in the water body were selected as the numerical experimental data for this paper. The data period is from December 1991 to December 2021, with a total of 458 samples. The commonly used statistical discriminant 3σ criterion is adopted to effectively discriminate and eliminate outliers. An empty string labeled 'unknown' was used to fill gaps. In this paper, the sample data are divided

Table 2 | Corresponding versions of software

No.	Software	Versions
1	JDK	jdk 1.8.0_152
2	Spark	spark-2.0.0-bin-hadoop2.6

into the following two parts: training samples and testing samples. The training sample is the water quality data from 1991 to 2019, and the test sample is the water quality data from 2020 to 2021.

4.2. Evaluation indicators

The prediction accuracy of the combined model is evaluated by the mean relative error (*MAPE*), root mean square error (*MSE*) and mean absolute error (*MAE*) between the predicted and true observed values. The three evaluation indicators are calculated as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{7}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{8}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{9}$$

where N represents the number of test samples, $\hat{y}_i$ represents the predicted value at the ith time point, and y_i represents the true observed value at the same time point.

4.3. Comparison of water quality prediction results

Five models, LIBSVM, Cascade SVM, EMD–Cascade SVM, EEMD–Cascade SVM and EEMD–Spark–Cascade SVM, were used to predict DO, NH_3-N and P, and the results are shown in Table 3.

The experiments focus on comparing the advantages and disadvantages of different models in terms of prediction accuracy, training time and stability. Model accuracy is measured by the prediction accuracy on the test set. The training time includes the overall training time and the prediction time. Model stability is measured by changing the number of initial divisions and

Table 3 | Prediction results of three indicators by different models

Indicator/mg·l^{-1}	Models	MAPE/%	MSE	MAE	Time/s
DO	LIBSVM	6.03	0.5404	0.5698	44.3
	Cascade SVM	5.00	0.2714	0.379	75.5
	EMD–Cascade SVM	4.01	0.4257	0.5517	87.1
	EEMD–Cascade SVM	3.73	0.1813	0.3397	99.6
	EEMD–Spark–Cascade–Cascade SVM	3.72	0.1812	0.3381	53.7
NH_3-N	LIBSVM	12.40	5.9841	4.8362	47.1
	Cascade SVM	8.41	4.607	3.0632	76.4
	EMD–Cascade SVM	7.30	3.9223	2.2413	85.7
	EEMD–Cascade SVM	6.93	3.2133	1.8269	103.6
	EEMD–Spark–Cascade SVM	6.95	3.2198	1.8185	51.2
P	LIBSVM	8.91	0.3566	0.3761	43.6
	Cascade SVM	6.57	0.2567	0.2978	73.6
	EMD–Cascade SVM	5.52	0.1545	0.1843	87.2
	EEMD–Cascade SVM	4.99	0.1411	0.1635	96.1
	EEMD–Spark–Cascade SVM	4.93	0.1402	0.1631	47.1

observing the change of prediction accuracy of different models. Taking the DO metric as an example, a comparison of the prediction results of different models as well as the true results is shown in Figure 5. Prediction error of the proposed model for DO index sequence is shown in Figure 6.

From Table 3 and Figure 5, it can be seen that the proposed combined prediction model EEMD–Spark–Cascade SVM obtains the highest prediction accuracy (by comparing MSE, MAPE and MAE indicators). This is because EEMD can completely uncover the intrinsic connection between water quality data, making the original data series smoother after the noise processing. EEMD solves the problem of modal confounding in the process of EMD method, and can more clearly show the fluctuation trend of the original series, thus effectively improving the performance of the model in terms of prediction accuracy. In addition, compared with the EEMD–Cascade SVM model, the combined prediction model EEMD–Spark–Cascade

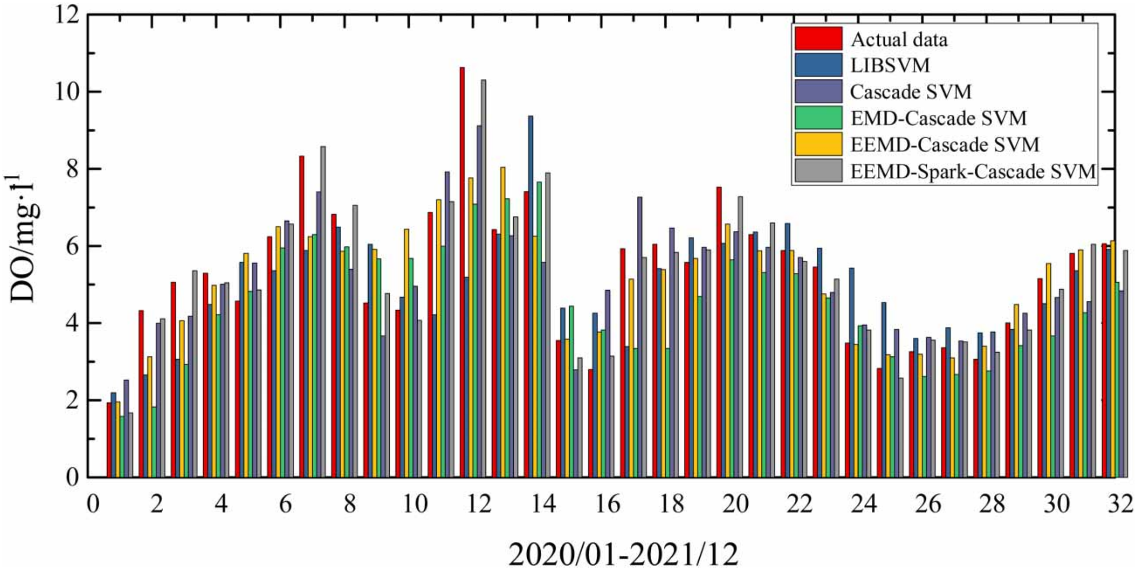


Figure 5 | Prediction comparison of each model for the DO index sequence.

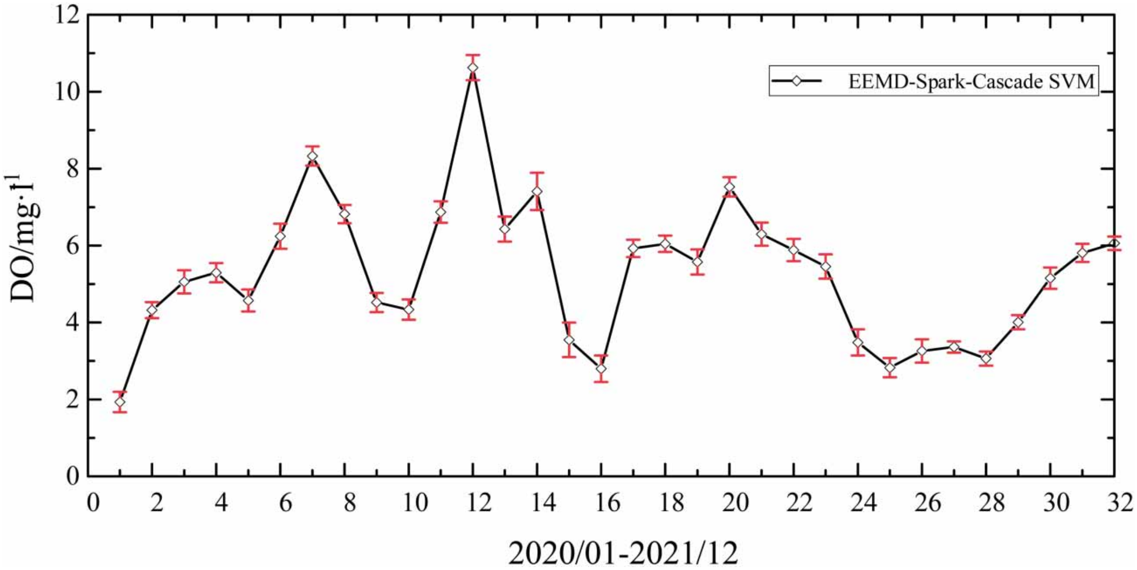


Figure 6 | Prediction error of the proposed model for the DO index sequence.

SVM has significantly reduced the training and prediction time, whereas the prediction accuracy remains unchanged. It can be seen that the parallel prediction time is greatly reduced compared to the standalone prediction time by a factor of about one. Moreover, with the increase of cluster size and parallelism, the time of parallel prediction achieved with the Spark platform can be reduced even more. Therefore, we can conclude that the EEMD–Spark–Cascade SVM water quality prediction

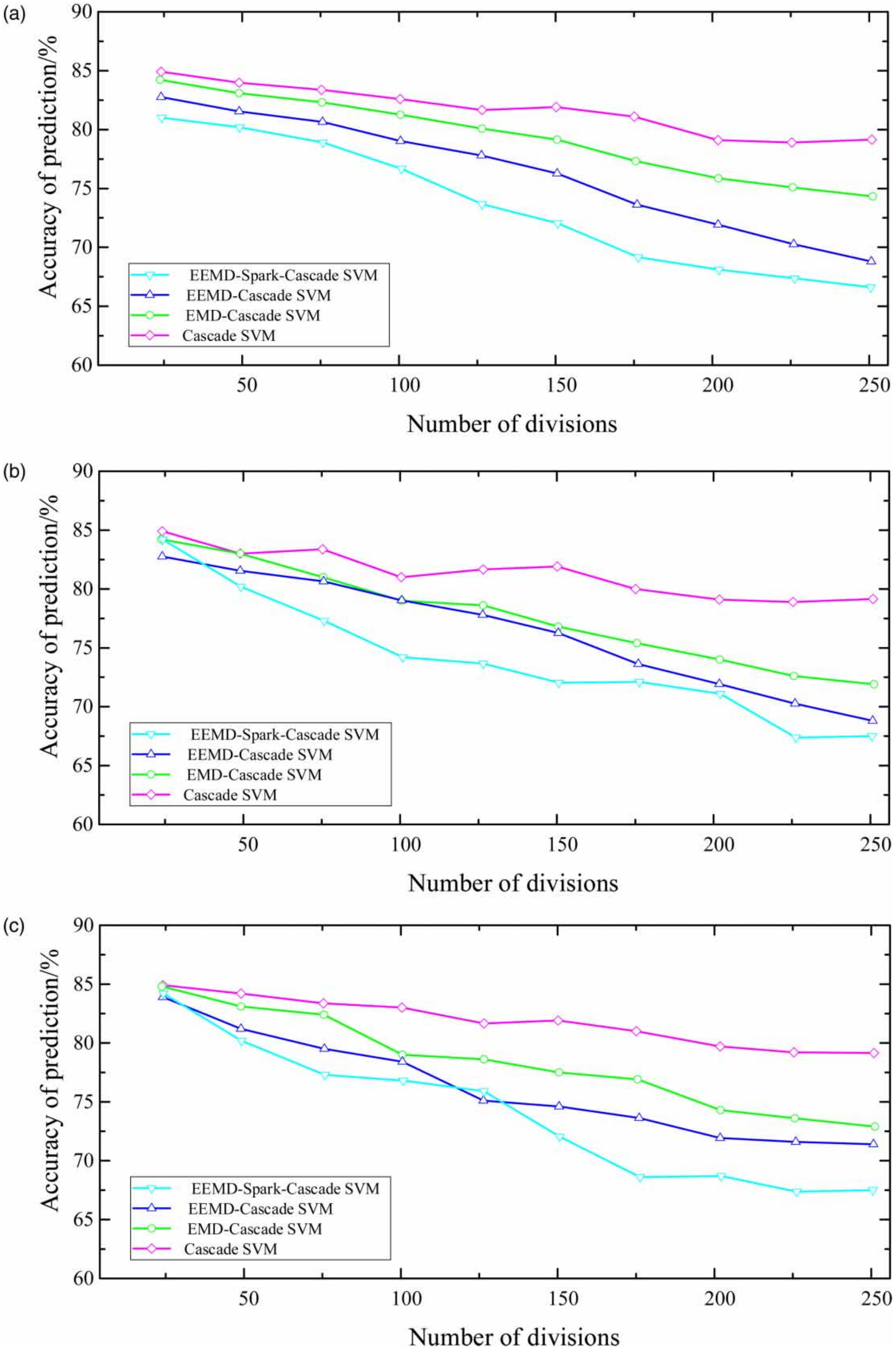


Figure 7 | Prediction accuracy under different division numbers: (a) DO; (b) NH_3-N and (c) P.

model is statistically superior to the other benchmark models considered. The proposed EEMD does not use the probability distribution of independent residuals to analyze water quality data, but instead uses a combination of residuals as shown in Equation (1).

To observe the effect of the number of divisions, the prediction accuracy of four models, Cascade SVM, EMD–Cascade SVM, EEMD–Cascade SVM and EEMD–Spark–Cascade SVM with different number of divisions, was analyzed. The experimental results are shown in Figure 7.

It can be seen from Figure 7 that the prediction accuracy of all four models, Cascade SVM, EMD–Cascade SVM, EEMD–Cascade SVM, and EEMD–Spark–Cascade SVM, decreases on the test set as the number of initial divisions continues to increase. This is because as the number of initial divisions increases, each subset generated by the random initial division becomes more different from the original data distribution, and some of the global SVs may be filtered out after the first layer of training. However, from the decreasing trend of the prediction accuracy of the four models, the decreasing speed of EEMD–Spark–Cascade SVM is significantly slower and less fluctuating compared with the other three models. In other words, the EEMD–Spark–Cascade SVM declines slowly with the increase of the number of divisions, and the decline rate is smoother and more stable.

5. CONCLUSIONS

A combined water quality prediction model is proposed in this paper based on EEMD and Cascade SVM. Considering the characteristics of water quality data such as nonlinearity and its instability, EEMD is introduced in the water quality prediction of the data for the processing of time and frequency, so as to reduce the instability of time series data, and more effectively improve the accuracy of the prediction results. Considering the low efficiency of water quality prediction on large-scale data, the parallelized Cascade SVM based on the Spark is used for each component obtained from the decomposition for prediction. The monthly data of DO, NH_3-N and P in water bodies were selected for experimental analysis, and the results showed that the combined water quality prediction model proposed in this paper has a higher accuracy and less prediction time compared with other prediction models. The shortcoming of this study is that the prediction accuracy of the model will not decrease with the increase in the number of initial divisions, and subsequently will use the quadratic distribution of the subset after the initial division to try to solve this problem.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Avila, R., Horn, B. & Moriarty, E. 2017 Evaluating statistical model performance in water quality prediction. *Journal of Environmental Management* **206**, 910–919.

Cao, Y., Yin, K. L. & Zhou, C. 2020 Establishment of landslide groundwater level prediction model based on GA-SVM and influencing factor analysis. *Sensors* **20** (3), 845–858.

Chen, D., Lin, J. & Li, Y. 2018 Modified complementary ensemble empirical mode decomposition and intrinsic mode functions evaluation index for high-speed train gearbox fault diagnosis. *Journal of Sound & Vibration* **424**, 192–207.

Cheng, W. C. & Jhan, D. M. 2013 Triaxial accelerometer-based fall detection method using a self-constructing cascade-AdaBoost-SVM classifier. *IEEE Journal of Biomedical and Health Informatics* **17** (2), 411–419.

Deng, T., Chau, K. W. & Duan, H. F. 2021 Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management* **284** (2), 56–64.

Ding, X., Zhu, Q., Liu, L. & Zhai, A. 2019 Water quality safety prediction model for drinking water source areas in Three Gorges Reservoir and its application. *Ecological Indicators* **101** (6), 734–741.

Du, S., Liu, T., Li, G. & Huang, D. 2018 A fast and adaptive bi-dimensional empirical mode decomposition approach for filtering of workpiece surfaces using high definition metrology. *Journal of Manufacturing Systems* **46**, 247–263.

Haleem, A. A., Perumandla, N. & Naruta, Y. 2019 Preparation of nanostructured Ta 3 N 5 electrodes by alkaline hydrothermal treatment followed by NH-3 annealing and their improved water oxidation performance. *ACS Omega* **4** (4), 7815–7821.

Harun, S. N., Hanafiah, M. M. & Nizam, N. 2020 Water and soil physicochemical characteristics of different rice cultivation areas. *Applied Ecology and Environmental Research* **18** (5), 6775–6791.

Hrnjica, B. & Bonacci, O. 2019 Lake level prediction using feed forward and recurrent neural networks. *Water Resources Management* **33** (7), 2471–2484.

Hu, X., Wang, H. & Zhu, Y. 2019 Landscape characteristics affecting spatial patterns of water quality variation in a highly disturbed region. *International Journal of Environmental Research and Public Health* **16** (12), 2149–2161.

Huang, W., Meng, L., Zhang, W. & Zhang, D. 2017 In-Memory parallel processing of massive remotely sensed data using an Apache Spark on Hadoop YARN model. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* **10** (1), 3–19.

Katimon, A., Shahid, S. & Mohsenipour, M. 2018 Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia. *Sustainable Water Resources Management* **4** (4), 991–998.

Khadr, M. & Elshemy, M. 2017 Data-driven modeling for water quality prediction case study: the drains system associated with Manzala Lake, Egypt. *Ain Shams Engineering Journal* **8** (4), 549–557.

Lu, L., Yan, J., An, S. & Chen, W. 2018 Investigation of KDP crystal surface based on an improved bidimensional empirical mode decomposition method. *Applied Surface Science* **433** (3), 680–688.

Mazo, C., Alegre, E. & Trujillo, M. 2017 Classification of cardiovascular tissues using LBP based descriptors and a cascade SVM. *Computer Methods and Programs in Biomedicine* **11**, 1–10.

Park, S. H. & Koo, J. 2015 Application of transfer function ARIMA modeling for the sedimentation process on water treatment plant. *International Journal of Control & Automation* **8** (10), 129–138.

Qaderi, F. & Engineering, C. E. 2017 Prediction of the groundwater remediation costs for drinking use based on quality of water resource, using artificial neural network. *Journal of Cleaner Production* **161**, 840–849.

Rashid, M. M., Nayan, A. A., Simi, S. A., Saha, J., Rahman, M. O. & Kibria, M. G. 2021 IoT based smart water quality prediction for biofloc aquaculture. *International Journal of Advanced Computer Science and Applications* **12** (6), 56–65.

Searcy, R. T. & Boehm, A. B. 2021 A day at the beach: enabling coastal water quality prediction with high-frequency sampling and data-driven models. *Environmental Science and Technology* **55** (3), 1908–1918.

Setshedi, K. J., Mutingwende, N. & Ngqwala, N. P. 2021 The use of artificial neural networks to predict the physicochemical characteristics of water quality in three district municipalities, eastern cape province, South Africa. *International Journal of Environmental Research and Public Health* **18** (10), 5248.

Shi, Z. & Zou, Z. 2014 Applied study of ARIMA model based on wavelet analysis on water quality prediction. *Chinese Journal of Environmental Engineering* **8** (10), 4550–4554.

Shihab, A. S. & Al-Tayyar, T. 2019 Prediction water quality in a selected sites at a stretch of Tigris river. *Journal of Engineering Science and Technology* **4** (3), 98–109.

Singha, S., Pasupuleti, S., Singha, S. S., Singh, R. & Kumar, S. 2021 Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **276** (4), 141–150.

Wang, L., Xiao, F. & Huang, C. 2019 Adaptive topology control with link quality prediction for underwater sensor networks. *Ad-hoc & Sensor Wireless Networks* **43** (3–4), 179–212.

Wu, F. & Meng, G. 2006 Compound rub malfunctions feature extraction based on full-spectrum cascade analysis and SVM. *Mechanical Systems & Signal Processing* **20** (8), 2007–2021.

Yi, H. S., Lee, B., Park, S., Kwak, K. C. & An, K. G. 2019 Prediction of short-term algal bloom using the M5P model-tree and extreme learning machine. *Environmental Engineering Research* **24** (3), 404–411.

Zhai, A., Hou, B., Huang, D. & Ding, X. 2021 Hazardous chemical accident prediction for drinking water sources in Three Gorges Reservoir. *Journal of Cleaner Production* **296** (6), 126529.

Zhang, L., Zhang, G. X. & Li, R. R. 2016 Water quality analysis and prediction using hybrid time series and neural network models. *Journal of Agricultural Science and Technology* **18** (4), 975–983.

Zhang, J., Ye, Z. & Zheng, K. 2021 A parallel computing approach to spatial neighboring analysis of large amounts of terrain data using spark. *Sensors* **21** (2), 365–372.

First received 20 March 2022; accepted in revised form 8 July 2022. Available online 20 July 2022