Water Supply

Water Supply Vol 22 No 1, 1107 doi: 10.2166/ws.2021.213

Estimating effluent turbidity in the drinking water flocculation process with an improved random forest model

Dongsheng Wang[a,b,*], Xiao Chang[a,b], Kaiwei Ma[a,b], Zhixuan Li[a,b] and Lianqing Deng[a,b]
[a] College of Automation & College of Artifical Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[b] Jiangsu Engineering Laboratory for Internet of Things and Intelligent Robots, Nanjing 210023, China
*Corresponding author. E-mail: wdsnjupt@163.com

DW, 0000-0003-4307-0992

ABSTRACT

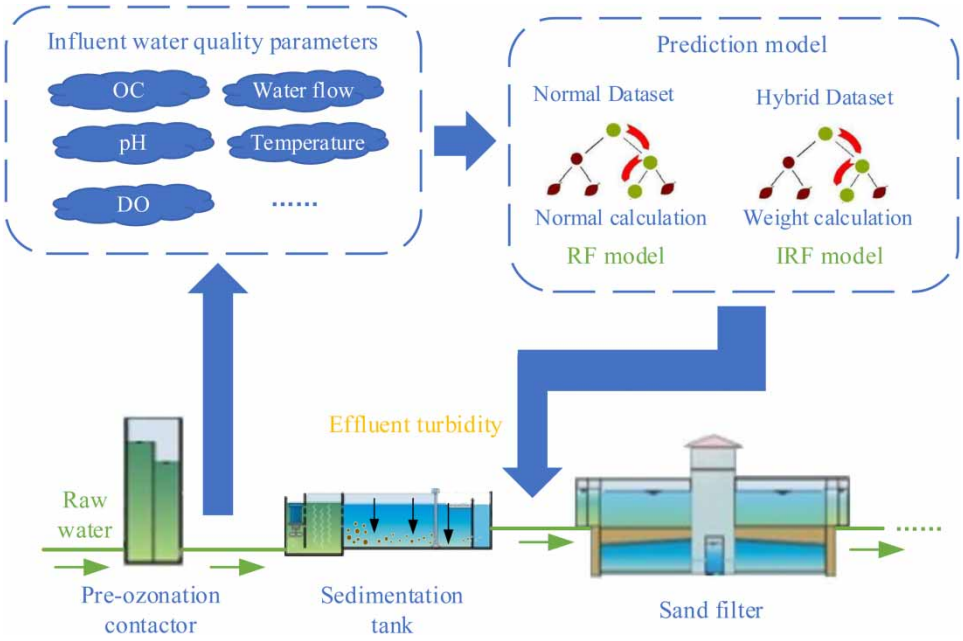
During drinking water treatment, the uncertain changes of raw water quality bring great difficulties to the control of flocculant dosage, especially because the feedback information based on the effluent turbidimeter of the sedimentation tank can only be obtained after a long time when the influent water quality changes due to the large lag characteristics of the flocculation process. Prediction of effluent turbidity of the sedimentation tank can effectively solve the aforementioned problem. Given that it is difficult for the ordinary random forest (RF) model to accurately predict the effluent turbidity of a sedimentation tank for complicated changes of raw water quality, an improved random forest (IRF) model composed of long-term and short-term parts is proposed, which can capture the periodicity and time-varying characteristics of influent water quality data. The experimental results show that the root mean square error and mean absolute percentage error of IRF model in Baiyangwan waterworks are improved 67.52% and 67.91% respectively, compared with those of the ordinary RF model. The proposed effluent turbidity predictions are also successfully developed in Xujiang waterworks and Xiangcheng waterworks of Suzhou, China. This research provides an effective method for real-time prediction of the effluent turbidity of sedimentation tank according to the influent water quality data.

Key words: drinking water, effluent turbidity, estimation model, flocculation process, improved random forest

HIGHLIGHTS

- A turbidity prediction method according to the influent water quality can provide real-time feedback information for flocculant dosing control.
- The IRF model composed of long-term and short-term parts captures the periodicity and time-varying characteristics of influent water quality.
- The IRF model has been experimented successfully in three water treatment plants with different raw water sources.

GRAPHICAL ABSTRACT



INTRODUCTION

During drinking water treatment, the flocculation process is a very important link (Di Marcantonio *et al.* 2020; Kim *et al.* 2020; Shao *et al.* 2020), which can form colloidal substances in water to absorb suspended impurities (Soros *et al.* 2019; Malkoske *et al.* 2020). The ideal flocculation process should dose appropriate flocculant according to the influent water quality index (Chua *et al.* 2020; Mohtar *et al.* 2020). The flocculant dosage directly determines the water purification effect (Xia *et al.* 2018). Within the flocculation process, effluent turbidity is a key evaluation indicator for effect (Liu & Ratnaweera 2016; Mucha & Kulakowski 2016; Vaananen *et al.* 2017; Melo *et al.* 2019). Generally, it takes more than 2 hours for the flocculation process to get the turbidity data from the turbidimeter after flocculant dosing. This brings great uncertainty to the flocculation process and seriously affects the control accuracy of the flocculant dosage. In summary, it is necessary to obtain real-time effluent turbidity of the sedimentation tank outlet by a prediction method (Fabrika *et al.* 2018).

In recent years, many attempts have been made to find a turbidity prediction method. Zhu *et al.* (2020) used an NIR camera and image processing software with the corresponding color component to measure the water turbidity. Pesic *et al.* (2016) pointed out that an appropriate regression model can be used for short-term turbidity simulation. Pennock *et al.* (2018) proposed a hydrodynamic and land cover model with adaptive variables, which can predict the turbidity of sediment according to the dosage of coagulant. Mather & Johnson (2015) used a combined cluster analysis and classification tree approach to predict the stream turbidity of three rivers in the Mid-Atlantic region of the United States.

However, the periodicity and complexity of water turbidity prediction data bring great challenges to the successful application of the aforementioned turbidity prediction methods. At the same time, public health workers are also looking for new technology for real-time turbidity prediction (Maquin *et al.* 2017; McCurley & Jawitz 2017; Zhang *et al.* 2017; Bernardelli *et al.* 2020).

Currently, with the development of information technology, artificial intelligence (AI) technology has gradually entered the public vision and has been widely used in the field of water treatment (Bian & Wang 2020; Bowen *et al.* 2020; Dasgupta *et al.* 2020; Watanabe *et al.* 2020; Wu *et al.* 2020). Therefore, many AI techniques have been applied to turbidity prediction (Khairi *et al.* 2016; Baghalian & Ghodsian 2017; Daghbandan *et al.* 2019; Song & Zhang 2020; Zounemat-Kermani *et al.* 2020). Abba *et al.* (2019) proposed a neuro fuzzy ensemble technique to predict turbidity in water treatment plants. Nieto *et al.* (2014, 2020) proposed a new practical model for long-term prediction of turbidity based on support vector machine and

particle swarm optimization. Then, they established a new turbidity prediction model of sand filter water for micro irrigation using gaussian process regression.

These AI algorithms have good application effects in water treatment. However, the raw water quality of drinking water treatment has a certain periodicity and seasonality. Here, we use the improved random forest (IRF) to solve this problem (Baral & Haq 2020; Liang *et al.* 2020; Liu *et al.* 2020). Before that, the random forest (RF) algorithm has been successfully applied for solving regression and classification problems in many applications (Mohammed *et al.* 2017; Li *et al.* 2020). It is suitable for demonstrating the nonlinear effect of variables, and it can model complex interactions among variables (Chen *et al.* 2020; Chencho *et al.* 2020; Kou *et al.* 2020; Zhang & Yang 2020). However, the common RF has difficulty coping with seasonal and periodic changes in influent water quality (Peng *et al.* 2020; Yang *et al.* 2020). Therefore, the IRF model is developed, which is a hybrid model consisting of long-term parts and short-term parts.

The main contribution of this work is the development of a practical and advanced soft sensor modeling method. It can real-time predict the turbidity at the outlet of sedimentation tank and provide the feedback information for flocculant dosing. The authors studied the characteristics of influent water quality of flocculation process and then propose a hybrid model based on the IRF. One novelty of this work lies in the fact that the IRF model can cope with seasonal and random changes in influent water quality. The other novelty of this work is that within the IRF model, the proportion of the long-term part and short-term part is adaptively updated according to the prediction accuracy.

DATA AND METHODS

Study area

The study area includes three drinking water treatment plants: Baiyangwan, Xiangcheng and Xujiang in Suzhou, China. The region is in the economically developed region of China. With the rapid development of the economy, the demand for drinking water in this area is increasing rapidly. Therefore, it is necessary to upgrade the existing drinking water treatment process. The whole process includes pre-chlorination, flocculation, sedimentation, sand filtration, ozonation, biological activated carbon and post-chlorination, as shown schematically in Figure 1. It can be seen that flocculation is the first key point to upgrade the process. On the one hand, it can remove most of the suspended impurities in the water and prepare for the subsequent treatment. On the other hand, it can regulate the quality of influent water and make the water quality stable.

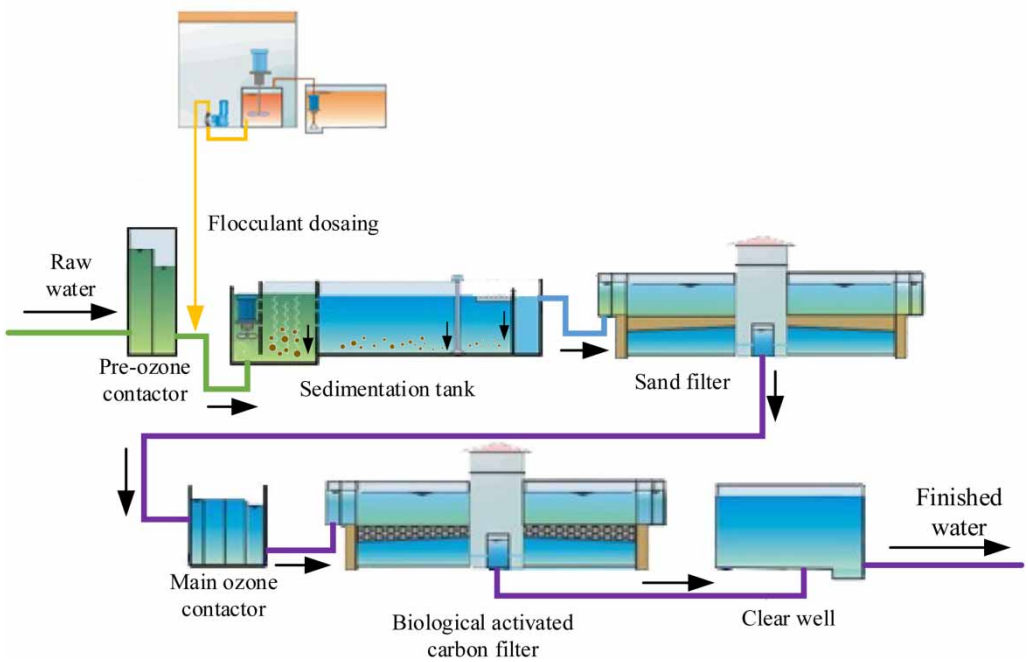


Figure 1 | Drinking water treatment process.

In these three water treatment plants, the most commonly used flocculant is alum. The main reason is that aluminum ion in alum forms colloidal adsorption particles in water and settles, which can quickly remove impurities in water. The specific addition method is as follows: the alum original solution is diluted with water to a 20% solution. Then, the flow ratio control method is used to add the agent to the inlet of the sedimentation tank, with the help of water for mixing. Finally, the flocculation is carried out in the sedimentation tank. Alum (aluminum sulfate at a concentration of 15 Baume degrees) was purchased from Suzhou Kang Shuo Chemical Co. Ltd (Jiangsu, China). Here, 15 Baume degrees refer to the specific gravity of aluminum sulfate in alum solution.

Monitoring data

The prediction of turbidity needs to monitor the water quality data. The amount of annual data is too large, and the data in July and August are highly representative. Therefore, the influent water quality data of July and August from 2015 to 2019 used in the study were measured by the study area. Because the water treatment plants only collect common influencing factors, the water quality variables for modeling include seven water quality variables: dissolved oxygen (DO), oxygen consumption (OC), pH, temperature, water flow, flocculant dosage and influent turbidity. Statistical analysis of daily water quality parameters is summarized in Table 1.

Data integration

Before data collection, we transfer the required data from the detection equipment. The equipment adopts structured programming to form the control system program of data acquisition and realize the linkage of the whole program.

Promoting the quality of measurement data is of great significance to any modeling method. So, before further analysis, a data preprocessing step is implemented. As the turbidity of effluent water is reflected two hours after alum dosing, the turbidity of effluent water will be shifted forward for two hours. Therefore, the data are collected two hours in advance and filtered. In the original data, all variable missing values are deleted. Different evaluation indexes usually have different orders of magnitude and units, and this will influence the results of data analysis. In order to remove the influence of magnitude between data indexes and solve the problem of incomparability, the data are normalized. Normalization is to map all the data to the range of 0–1, which is convenient and fast. Then, the turbidity is obtained by inverse normalization of the prediction results.

RF model

RF was proposed by Leo Breiman, who was enlightened by the early work of Amit and Geman (Cutler *et al.* 2012). RF is an extension of Breiman's bagging concept and has developed into a competitor to enhance packaging. RF can be used for categorical response variables (called 'classification') or continuous responses (called 'regression'). Similarly, predicted parameters can be categorical or continuous parameters. The structure of RF is shown in Figure 2.

As the name suggests, RF is based on the set of trees, and each tree depends on the set of random variables. More regularly, the p-dimensional random vector $X = (X_1, \ldots, X_P)^T$ expresses as the input variable or output variable, and the random variable Y expresses as the response. Let's assume that the unknown joint distribution is $P_{XY}(X, Y)$. The purpose of this assumption is to look for a prediction function $f(X)$ used to predict Y. The prediction function is determined by the loss function $L(Y, f(X))$ and is defined as the minimum loss of $E_{XY}(L(Y, f(X)))$. In addition, the subscript expresses as the expected value of the joint distribution of X and Y.

Table 1 | Influent water quality of the study area in July and August 2015–2019

	Max	Min	Average
DO (mg/L)	20.07	0.75	6.49
OC (mg/L)	6.19	2.84	4.03
pH	8.93	7.15	8.04
Temperature (°C)	35.76	25.68	30.66
Water flow (m^3)	3,540.00	1,104.88	3,108.75
Flocculant dosage (mg/L)	85.10	7.04	36.74
Influent turbidity (NTU)	121.42	4.48	24.73

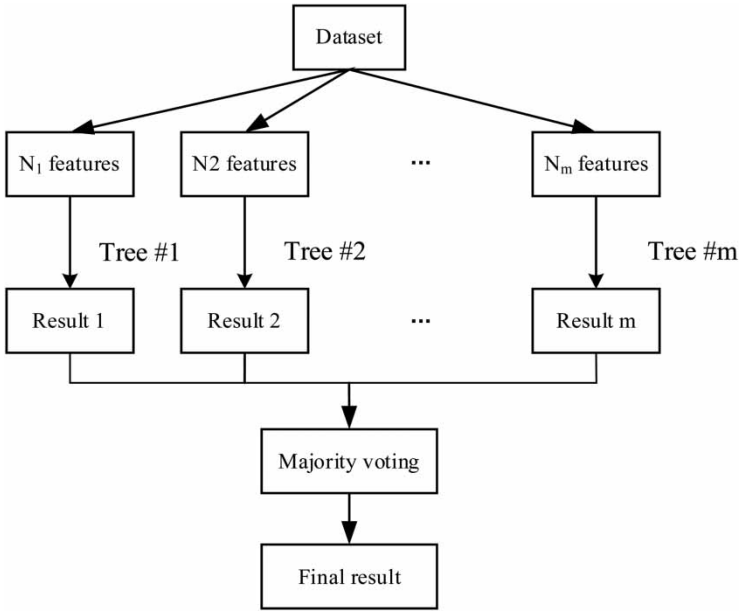


Figure 2 | Structure of the RF.

Intuitively, $L(Y, f(X))$ is a measure of the closeness between $f(X)$ and Y. It penalizes the value of $f(X)$ far away from Y. The representative choice of L is the square error loss $L(Y, f(X)) = (Y\text{-}f(X))^2$ for regression and zero-one loss for classification:

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0, Y = f(X) \\ 1, otherwise \end{cases} \tag{1}$$

It is proved that the conditional expectation is given for the minimization of square error loss $E_{XY}(L(Y, f(X)))$:

$$f(X) = E(Y|X = x) \tag{2}$$

Otherwise, it is called the regression method. In the terms of classification, if β is used to represent the set of possible Y, then the zero-one $E_{XY}(L(Y, f(X)))$ is minimized:

$$f(x) = arg\max_{y \in \beta} p(Y = y|X = x) \tag{3}$$

Otherwise, it is called the Bayes rule.

According to the so-called 'base learners' $h_1(x), \ldots, h_J(x)$ to construct the set f, and then combine these basic learners to get the 'global predictor' $f(x)$. In the regression, the average of the base learners is:

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x) \tag{4}$$

In the classification, $f(x)$ is the most frequently predicted class ('voting'):

$$f(x) = arg\max_{y \in \beta} \sum_{j=1}^{J} I(y = h_j(x)) \tag{5}$$

In RF, the j-th base learner is a tree, expressed as $h_j(X, \theta_j)$, where θ_j is a set of random variables, and θ_j for $j = 1, \ldots, J$ is independent. Although the definition of RF is very general, RF is always implemented in a specific way. To find out the RF algorithm, it is significant to master the basic knowledge of the types of trees used as base learners.

RF is suitable for demonstrating the nonlinear effect of variables and can simulate the complex interaction between variables. This is consistent with the prediction of effluent turbidity in drinking water flocculation.

IRF model

The common RF has difficulty coping with seasonal and periodic changes in influent water quality. Therefore, an IRF model with a hybrid model is proposed. The hybrid model of turbidity prediction is divided into long-term and short-term models. The long-term data are collected in July and August 2015–2019, while the short-term data are collected every seven days. IRF model compares the actual turbidity with the short-term part forecast turbidity and the long-term part forecast turbidity to determine which part is closer to the actual turbidity. Then, the weight is adaptively obtained according to the degree of deviation. The structure of IRF is shown in Figure 3.

The steps of adaptive weighting are as follows:

(1) Calculate the deviation σ between the effluent turbidity of the sedimentation tank and the set value α at the current time. If the deviation is greater than a certain threshold φ, it indicates that the effluent turbidity at the corresponding time is unreasonable, so start step (2). Otherwise, the effluent turbidity is reasonable, and the current weight remains unchanged.

(2) When the effluent turbidity of the sedimentation tank is greater than the set value, it indicates that the actual flocculant dosage at the corresponding time is too small. At this time, the current long-term and short-term model weights are adjusted according to the principle of increasing the flocculant dosage. This reduces the weight of the part, of which the output is larger among the long-term part and short-term part (adjustment cycle is 1 hour/time). In contrast, it indicates that the actual flocculant dosage at the corresponding time is too large, and the current long-term and short-term model weights are adjusted according to the principle of reducing the flocculant dosage. This increases the weight of

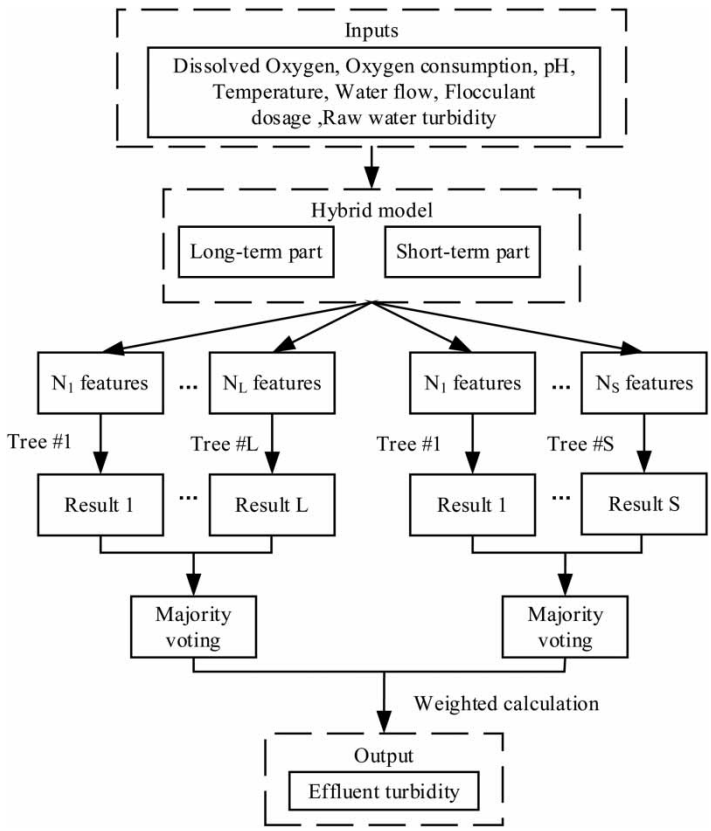


Figure 3 | Structure of the IRF.

the part, of which the output is larger among the long-term part and short-term part. Then, the effluent turbidity of the hybrid model is calculated according to the weight.

The weighted formula of the hybrid model is as follows:

$$\begin{cases} y = ay_L + by_S \\ a + b = 1 \end{cases} \tag{6}$$

where y is the prediction value of the hybrid model, y_L and y_S are the prediction values of the long-term part and the short-term part, respectively, and a and b are the weights of the long-term part and the short-term part, respectively.

The IRF model divides the data source into long-term parts and short-term parts for weighted calculation, which reasonably optimizes the data structure and caters to the periodicity and time-varying nature of water quality data in the algorithm structure.

Assessment of model performance

To verify the effect of the model, the root mean square error (RMSE) and mean absolute percentage error (MAPE) are set as the evaluation indexes. Their expressions are as follows:

$$RMSE = \sqrt{\frac{\sum_{i,l=1}^{n} (y_l - y_i)^2}{n}} \tag{7}$$

$$MAPE = \frac{1}{n} \sum_{i,l=1}^{n} \left| 100 \times \frac{y_i - y_l}{y_i} \right| \tag{8}$$

where y_i and y_l are the measured value and predicted value of the output variable, respectively. n represents the number of test samples. These predictive performance indicators provide different interpretations for the modeling results. RMSE represents the prediction accuracy of the prediction model, and MAPE gives the average ratio of the error to the measured value.

The overall procedure using the prediction models is listed below:

I. Collect the input and output data of flocculation process of drinking water treatment for prediction model training;
II. Implement the data preprocessing step consisting of data cleaning, data transformation and data reduction;
III. The RF and IRF models are trained by using the treated influent water quality data;
IV. Use root mean square error (RMSE) and mean absolute percentage error (MAPE) between the forecast data and the observation data to compare the forecast results, which are shown in (7) and (8), respectively.

RESULTS

Descriptive statistics

The statistics of the water quality variables are summarized in Table 1. During the whole study period, the average effluent turbidity of the Suzhou region was 1.07, and the effluent turbidity of the three water treatment plants had little difference. The average effluent turbidity of the Baiyangwan water treatment plant, Xiangcheng water treatment plant and Xujiang water treatment plant are 0.99, 0.97 and 1.12, respectively. In addition, due to the different geographical locations of each water plant, the set value of effluent turbidity is also different. Although the data of different waterworks are different, the difference is not large. So the data of the Baiyangwan water treatment plant can be selected for RF and IRF comparative tests.

Estimation modeling and validation

In this study, an RF model is first constructed, with seven variables (dissolved oxygen, oxygen consumption, pH, temperature, water flow, flocculant dosage and influent turbidity) as input variables and effluent turbidity as the output variable for prediction. Then, the same prediction is made using the IRF model.

In the prediction of effluent turbidity, RF uses the regression method. In the training part, the RF model collects 15 different sub-training data sets from the data set using bootstrap sampling, and trains 15 different decision trees in turn. In the

prediction part, the RF model averages the prediction results of 15 internal decision trees to obtain the final result. The prediction results of effluent turbidity using RF are shown in Figure 4.

The hybrid model based on the IRF in this study is divided into two parts: the long-term part and short-term part. The final turbidity is obtained by adaptive weighted calculation of the predicted value of the long-term part and short-term part. The certain threshold φ of the hybrid model is 0.3. In the training phase, the long-term part of the IRF model collects 20 different sub training datasets for training, and the short-term part collects eight different sub training datasets for training. In addition, the weight of the hybrid model is adaptive. Among them, the long-term part can show seasonal changes, the short-term part can show random changes, and the establishment of the hybrid model significantly improves the accuracy of turbidity prediction. The prediction results of effluent turbidity using IRF are shown in Figure 5.

In Figures 4 and 5, the uncertainty of turbidity prediction is analyzed by dividing the 95% confidence interval. The results show that the confidence interval of IRF can contain 98.39% of the measured samples. RF can only contain 80.65%. Especially in the test phase, the RMSE index of IRF is 0.0293, and the RMSE index of RF is 0.0902. It can be seen from the above results that the effect of the IRF model in Baiyangwan waterworks is better than that of the RF model. Therefore, IRF is better than RF in predicting effluent turbidity.

Estimated turbidity over Suzhou region

The IRF model is also used to predict the effluent turbidity of the sedimentation tank in the Xiangcheng water plant and Xujiang water plant. The prediction results are shown in Figures 6 and 7.

The IRF model was used to predict the effluent turbidity of three water treatment plants in the study area. These plants have their own set values of effluent turbidity, which are 1.0 for Baiyangwan water treatment plant and Xiangcheng water treatment plant, and 1.2 for Xujiang water treatment plant. Figure 8 shows the application effect of the IRF model in three water treatment plants in the Suzhou region through a residual diagram. The residual value of effluent turbidity of the sedimentation tank is less than 0.15, which indicates that the application effect of the IRF model is good.

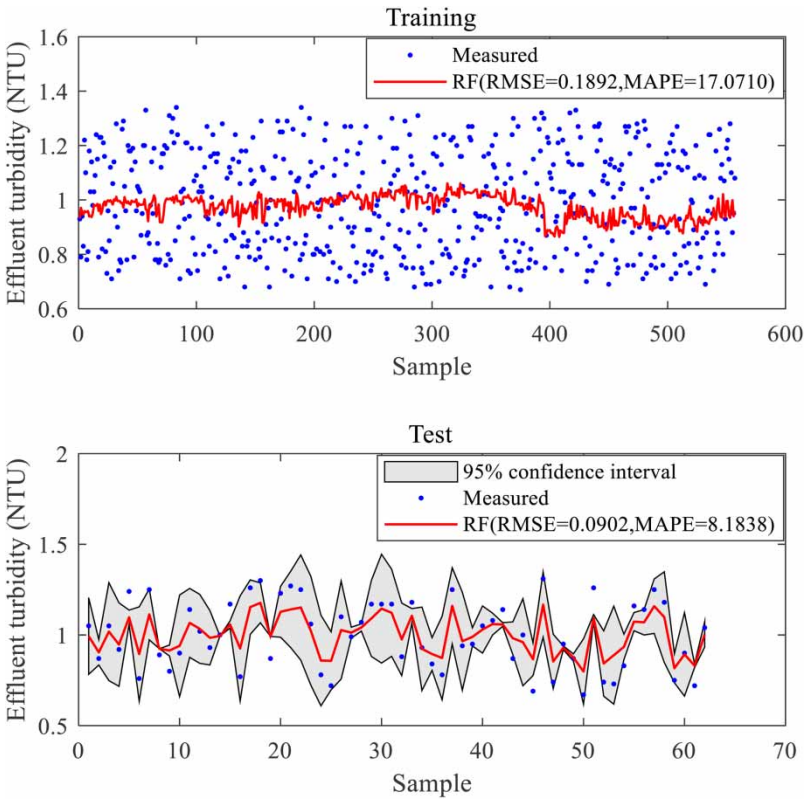


Figure 4 | Prediction results of effluent turbidity using RF in Baiyangwan water treatment plant.

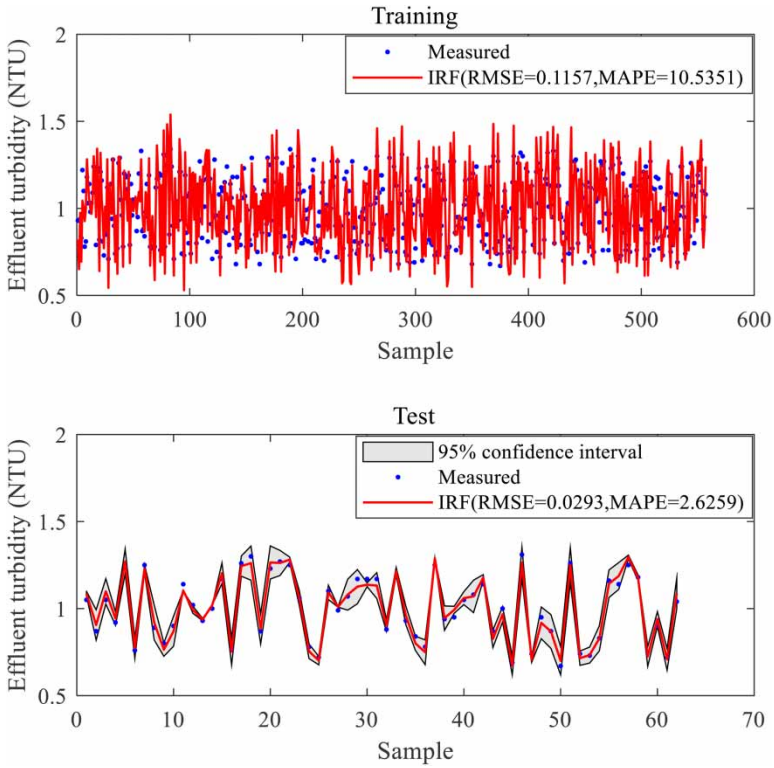


Figure 5 | Prediction results of effluent turbidity using IRF in Baiyangwan water treatment plant.

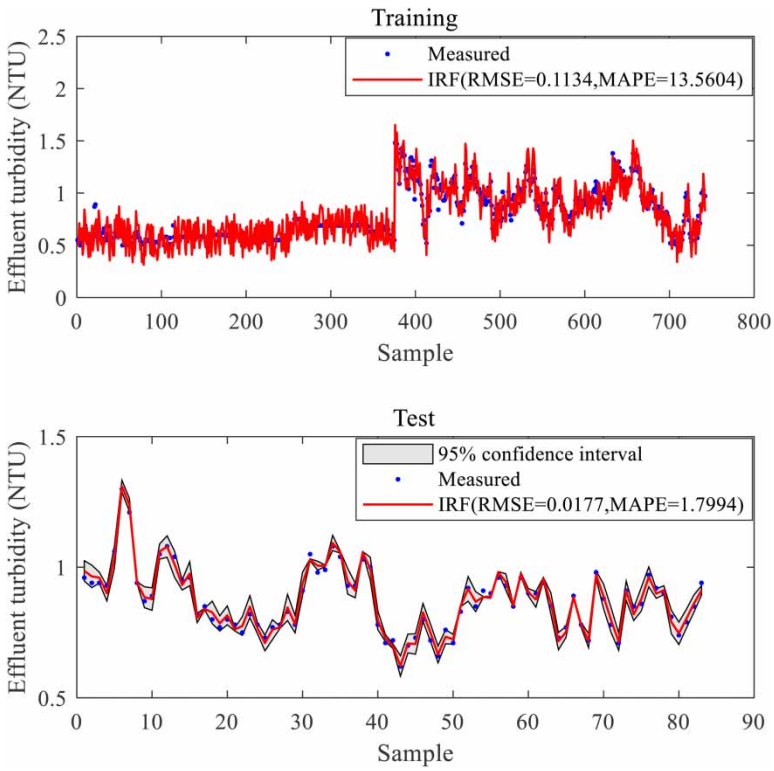


Figure 6 | Prediction results of effluent turbidity using IRF in Xiangcheng water treatment plant.

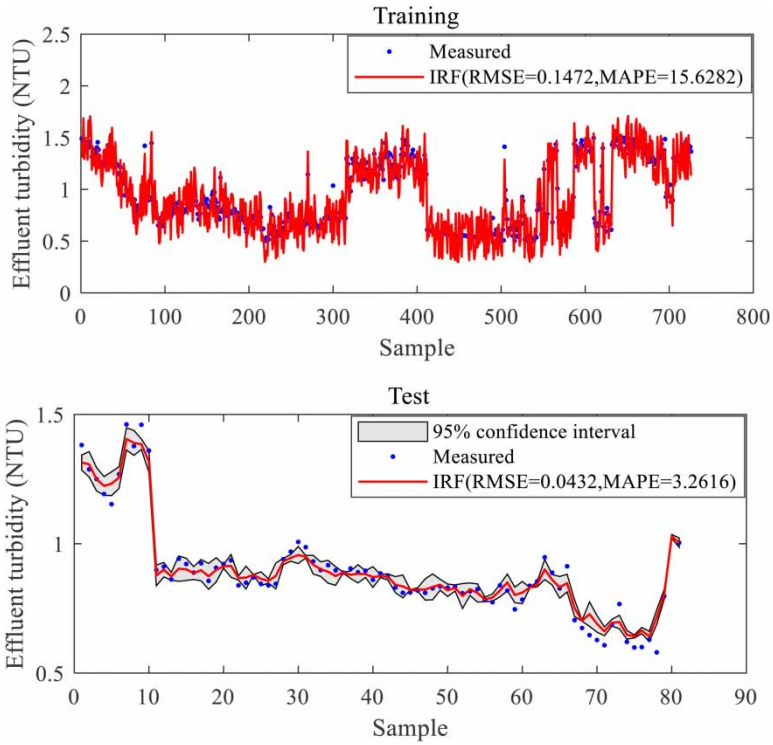


Figure 7 | Prediction results of effluent turbidity using IRF in Xujiang water treatment plant.

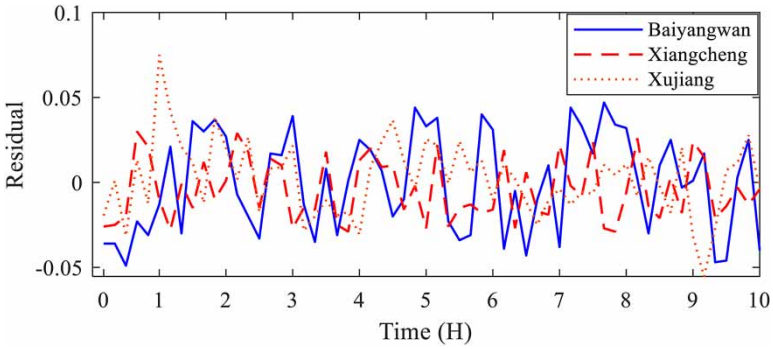


Figure 8 | The residual diagram of effluent turbidity in three water treatment plants in Suzhou region.

In addition, weight calculation is used in the process of effluent turbidity prediction of the three water treatment plants in the Suzhou region. The time taken by the three water treatment plants was 10 hours, and the weight was updated every hour. The weight ratio of each time period in Figure 8 is shown in Table 2. Table 2 clearly shows the weight ratio changes of the three water treatment plants in the 10 hours' sampling time.

DISCUSSION

In this study, an artificial intelligence method based on the IRF model is used to predict the effluent turbidity of sedimentation tanks for the purpose of control of flocculant dosage. The traditional flocculant dosage is usually adjusted by feedback control based on the turbidimeter information. The proposed method can provide real-time prediction of the effluent turbidity according to the raw water quality instead of waiting a very long time for the turbidity meter information.

In the application of a non-artificial intelligence method in the flocculant dosing process (Chua *et al.* 2020; Mohtar *et al.* 2020), it is often necessary to model separately according to the situation of water plants, and this model cannot be directly applied to other water plants. The IRF model can adjust parameters automatically and can be directly applied to multiple

Table 2 | The hourly weight ratio of the different water treatment plants

Items	Time (H)	Weight ratio of long-term model and short-term model
Baiyangwan	0–1	0.55:0.45
	1–2	0.60:0.40
	2–3	0.58:0.42
	3–4	0.58:0.42
	4–5	0.60:0.40
	5–6	0.61:0.39
	6–7	0.60:0.40
	7–8	0.58:0.42
	8–9	0.59:0.41
	9–10	0.58:0.42
Xiangcheng	0–1	0.66:0.34
	1–2	0.62:0.38
	2–3	0.48:0.52
	3–4	0.51:0.49
	4–5	0.60:0.40
	5–6	0.58:0.42
	6–7	0.61:0.39
	7–8	0.60:0.40
	8–9	0.60:0.40
	9–10	0.60:0.40
Xujiang	0–1	0.46:0.54
	1–2	0.55:0.45
	2–3	0.58:0.42
	3–4	0.57:0.43
	4–5	0.55:0.45
	5–6	0.56:0.44
	6–7	0.55:0.45
	7–8	0.55:0.45
	8–9	0.58:0.42
	9–10	0.54:0.46

waterworks. Compared with the traditional method, the model has strong adaptability, can adjust parameters adaptively and is more convenient.

The IRF model has the unique advantage of dividing the data into long-term data and short-term data, and using adaptive weight updating. It can be seen from Table 2 that most weight ratios are biased towards long-term data, which indicates that the periodicity of water quality data is dominant in most cases. In addition, occasionally, the weight ratio tends to short-term data, which indicates that the water quality data has a certain randomness. The evaluating indicators also show that the IRF model has high accuracy in predicting effluent turbidity; for example, the RMSE and MAPE of the IRF model are better than that of the RF model.

CONCLUSION

Here, we propose an IRF model to predict the effluent turbidity of sedimentation tanks in Suzhou water treatment plants by using a hybrid model with long-term and short-term parts. In order to show that the IRF model is more accurate, three water treatment plants were selected in the study area for experimental application.

We also compare the performance of the RF model and the IRF model, the RMSE and MAPE of the IRF model in the Baiyangwan waterworks are improved 67.52% and 67.91%, compared with that of ordinary RF model, respectively. Tests in other water treatment plants also show that the IRF model has a strong prediction ability for effluent turbidity of the sedimentation tank. At the same time, it can be directly applied in different waterworks, which shows that the IRF model has high flexibility and adaptability, and is more attractive for modeling of a drinking water flocculation process with complex characteristics. Therefore, for water quality data with time-varying and periodic characteristics, the IRF model has better prediction performance.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (51708299), Science and Technology Project of Water Conservancy of Jiangsu Province (2020056), Major Science and Technology Program for Water Pollution Control and Treatment (2012ZX07403-001) and the NUPTSF (NY220140).

DATA AVAILABILITY STATEMENT

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

REFERENCES

Abba, S. I., Abdulkadir, R. A., Gaya, M. S., Saleh, M. A., Esmaili, P. & Jibril, M. B. 2019 Neuro-fuzzy ensemble techniques for the prediction of turbidity in water treatment plant. In *2019 2nd International Conference of the Ieee Nigeria Computer Chapter (Nigeriacomputconf)*. pp. 117–122.

Baghalian, S. & Ghodsian, M. 2017 Experimental analysis and prediction of velocity profiles of turbidity current in a channel with abrupt slope using artificial neural network. *Journal of the Brazilian Society of Mechanical Sciences and Engineering* **39** (11), 4503–4517.

Baral, P. & Haq, M. A. 2020 Spatial prediction of permafrost occurrence in Sikkim Himalayas using logistic regression, random forests, support vector machines and neural networks. *Geomorphology* **371** (15), 107331.

Bernardelli, A., Marsili-Libelli, S., Manzini, A., Stancari, S., Tardini, G., Montanari, D., Anceschi, G., Gelli, P. & Venier, S. 2020 Real-time model predictive control of a wastewater treatment plant based on machine learning. *Water Science and Technology* **81** (11), 2391–2400.

Bian, Q. & Wang, Z. D. 2020 Development of computer artificial intelligence ai and analysis of its hardware technology. *Basic & Clinical Pharmacology & Toxicology* **127**, 133–134.

Bowen, T. J., Stephens, L., Vance, M., Huang, Y. C., Fridman, D. & Nabhan, C. 2020 Novel artificial intelligence (AI)-based technology to improve oncology clinical trial fulfillment. *Journal of Clinical Oncology* **38** (15), 2052.

Chen, G. K., Zhang, X. C., Wu, Z. B., Su, J. H. & Cai, G. R. 2020 An efficient tea quality classification algorithm based on near infrared spectroscopy and random forest. *Journal of Food Process Engineering* **44** (1), e13604.

Chencho, L. J., Hao, H., Wang, R. H. & Li, L. 2020 Development and application of random forest technique for element level structural damage quantification. *Structural Control & Health Monitoring* **28** (3), e2678.

Chua, S. C., Chong, F. K., Malek, M. A., Ul Mustafa, M. R., Ismail, N., Sujarwo, W., Lim, J. W. & Ho, Y. C. 2020 Optimized use of ferric chloride and sesbania seed Gum (SSG) as sustainable coagulant aid for turbidity reduction in drinking water treatment. *Sustainability* **12** (6), 2273.

Cutler, A., Cutler, D. R. & Stevens, J. R. 2012 Random forests. In: *Ensemble Machine Learning: Methods and Applications* (Zhang, C. & Ma, Y., eds.). Springer, Boston, MA, US, pp. 157–175.

Daghbandan, A., Khalatbari, S. & Abbasi, M. M. 2019 Applying GMDH-type neural network for modeling and prediction of turbidity and free residual aluminium in drinking water. *Desalination and Water Treatment* **140**, 118–131.

Dasgupta, I., Saha, J., Venkatasubbu, P. & Ramasubramanian, P. 2020 AI crop predictor and weed detector using wireless technologies: a smart application for farmers. *Arabian Journal for Science and Engineering* **45** (12), 11115–11127.

Di Marcantonio, C., Bertelkamp, C., van Bel, N., Pronk, T. E., Timmers, P. H. A., van der Wielen, P. & Brunner, A. M. 2020 Organic micropollutant removal in full-scale rapid sand filters used for drinking water treatment in The Netherlands and Belgium. *Chemosphere* **260**, 127630.

Fabrika, M., Valent, P. & Scheer, L. 2018 Thinning trainer based on forest-growth model, virtual reality and computer-aided virtual environment. *Environmental Modelling & Software* **100**, 11–23.

Khairi, M. T. M., Ibrahim, S., Yunus, M. A. M., Faramarzi, M. & Yusuf, Z. 2016 Artificial neural network approach for predicting the water turbidity level using optical tomography. *Arabian Journal for Science and Engineering* **41** (9), 3369–3379.

Kim, K. Y., Ekpe, O. D., Lee, H. J. & Oh, J. E. 2020 Perfluoroalkyl substances and pharmaceuticals removal in full-scale drinking water treatment plants. *Journal of Hazardous Materials* **400**, 123235.

Kou, L., Liu, C., Cai, G. W., Zhou, J. N. & Yuan, Q. D. 2020 Data-driven design of fault diagnosis for three-phase PWM rectifier using random forests technique with transient synthetic features. *Iet Power Electronics* **13** (16), 3571–3579.

Li, Z. L., Wang, H. X., Zhang, Y. W. & Zhao, X. H. 2020 Random forest-based feature selection and detection method for drunk driving recognition. *International Journal of Distributed Sensor Networks* **16** (2), 1–13.

Liang, T. C., Sun, L. & Li, H. X. 2020 MODIS aerosol optical depth retrieval based on random forest approach. *Remote Sensing Letters* **12** (2), 179–189.

Liu, W. & Ratnaweera, H. 2016 Improvement of multi-parameter-based feed-forward coagulant dosing control systems with feed-back functionalities. *Water Science and Technology* **74** (2), 491–499.

Liu, J., Sun, S. Q., Tan, Z. L. & Liu, Y. 2020 Nondestructive detection of sunset yellow in cream based on near-infrared spectroscopy and interval random forest. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy* **242**, 118718.

Malkoske, T. A., Berube, P. R. & Andrews, R. C. 2020 Coagulation/flocculation prior to low pressure membranes in drinking water treatment: a review. *Environmental Science-Water Research & Technology* **6** (11), 2993–3023.

Maquin, M., Mouche, E., Mügler, C., Pierret, M.-C. & Viville, D. 2017 A soil column model for predicting the interaction between water table and evapotranspiration. *Water Resources Research* **53** (7), 5877–5898.

Mather, A. L. & Johnson, R. L. 2015 Event-based prediction of stream turbidity using a combined cluster analysis and classification tree approach. *Journal of Hydrology* **530**, 751–761.

McCurley, K. L. & Jawitz, J. W. 2017 Hyphenated hydrology: interdisciplinary evolution of water resource science. *Water Resources Research* **53** (4), 2972–2982.

Melo, L. D. V., da Costa, E. P., Pinto, C. C., Barroso, G. R. & Oliveira, S. C. 2019 Adequacy analysis of drinking water treatment technologies in regard to the parameter turbidity, considering the quality of natural waters treated by large-scale WTPs in Brazil. *Environmental Monitoring and Assessment* **191** (6), 1–12.

Mohammed, H., Hameed, I. A. & Seidu, R. 2017 Random forest tree for predicting fecal indicator organisms in drinking water supply. In *Proceedings of 4th International Conference on Behavioral, Economic Advance in Behavioral, Economic, Sociocultural Computing (Besc)*.

Mohtar, S. S., Sharuddin, S. S. N., Saman, N., Lye, J. W. P., Othman, N. S. & Mat, H. 2020 A simultaneous removal of ammonium and turbidity via an adsorptive coagulation for drinking water treatment process. *Environmental Science and Pollution Research* **27** (16), 20173–20186.

Mucha, Z. & Kulakowski, P. 2016 Turbidity measurements as a tool of monitoring and control of the SBR effluent at the small wastewater treatment plant – preliminary study. *Archives of Environmental Protection* **42** (3), 33–36.

Nieto, P. J. G., Garcia-Gonzalo, E., Fernandez, J. R. A. & Muniz, C. D. 2014 Hybrid PSO-SVM-based method for long-term forecasting of turbidity in the Nalon river basin: a case study in Northern Spain. *Ecological Engineering* **73**, 192–200.

Nieto, P. J. G., Garcia-Gonzalo, E., Puig-Bargues, J., Sole-Torres, C., Duran-Ros, M. & Arbat, G. 2020 A new predictive model for the outlet turbidity in micro-irrigation sand filters fed with effluents using Gaussian process regression. *Computers and Electronics in Agriculture* **170**, 105292.

Peng, C., Yang, M. Q., Zheng, Q. H., Zhang, J., Wang, D. Q., Yan, R. Y., Wang, J. X. & Li, B. J. 2020 A triple-thresholds pavement crack detection method leveraging random structured forest. *Construction and Building Materials* **263**, 120080.

Pennock, W. H., Weber-Shirk, M. L. & Lion, L. W. 2018 A hydrodynamic and surface coverage model capable of predicting settled effluent turbidity subsequent to hydraulic flocculation. *Environmental Engineering Science* **35** (12), 1273–1285.

Pesic, M., Vakanjac, V. R., Vakanjac, B. & Jovanov, K. 2016 Turbidity simulation for short-term predictions: case study of the Karst Spring Surdup (Bor, Serbia). *Comptes Rendus De L Academie Bulgare Des Sciences* **69** (9), 1183–1194.

Shao, Y., Zhou, X. H., Liu, X. W. & Wang, L. L. 2020 Pre-oxidization-induced change of physicochemical characteristics and removal behaviours in conventional drinking water treatment processes for polyethylene microplastics. *Rsc Advances* **10** (68), 41488–41494.

Song, C. Y. & Zhang, H. P. 2020 Study on turbidity prediction method of reservoirs based on long short term memory neural network. *Ecological Modelling* **432**, 109210.

Soros, A., Amburgey, J. E., Stauber, C. E., Sobsey, M. D. & Casanova, L. M. 2019 Turbidity reduction in drinking water by coagulation-flocculation with chitosan polymers. *Journal of Water and Health* **17** (2), 204–218.

Vaananen, J., Memet, S., Guenther, T., Lilja, M., Cimbritz, M. & Jansen, J. L. 2017 Automatic control of the effluent turbidity from a chemically enhanced primary treatment with microsieving. *Water Science and Technology* **76** (7), 1770–1780.

Watanabe, H., Fukunaga, N., Sanami, S., Kitasaka, H., Tsuzuki, Y., Kida, Y., Takeda, S., Kondo, F., Takeda, S. & Asada, Y. 2020 Comparison of pronuclear (Pn) number observations based on embryologist's experience and detection by artifical intelligence (Ai) trained with deep learning technology. *Fertility and Sterility* **114** (3), E75-E.

Wu, J., Ding, Z. R., Yu, H. X. & Yu, S. M. 2020 Application of AI technology in English writing teaching on coronavirus paper in multimedia environment. *Basic & Clinical Pharmacology & Toxicology* **127**, 240.

Xia, X., Lan, S. H., Li, X. D., Xie, Y. F., Liang, Y. J., Yan, P. H., Chen, Z. Y. & Xing, Y. X. 2018 Characterization and coagulation-flocculation performance of a composite flocculant in high-turbidity drinking water treatment. *Chemosphere* **206**, 701–708.

Yang, Q. Q., Yuan, Q. Q., Li, T. W. & Yue, L. W. 2020 Mapping PM2.5 concentration at high resolution using a cascade random forest based downscaling model: evaluation and application. *Journal of Cleaner Production* **277**, 123887.

Zhang, F. & Yang, X. J. 2020 Improving land cover classification in an urbanized coastal area by random forests: the role of variable selection. *Remote Sensing of Environment* **251**, 112105.

Zhang, J. Y., Du, C. C. & Feng, X. G. 2017 Research on a soft measurement model of sewage treatment based on a case-based reasoning approach. *Water Science and Technology* **76** (12), 3181–3189.

Zhu, Y. Y., Cao, P. P., Liu, S., Zheng, Y. & Huang, C. Q. 2020 Development of a new method for turbidity measurement using two NIR digital cameras. *Acs Omega* **5** (10), 5421–5428.

Zounemat-Kermani, M., Alizamir, M., Fadaee, M., Namboothiri, A. S. & Shiri, J. 2020 Online sequential extreme learning machine in river water quality (turbidity) prediction: a comparative study on different data mining approaches. *Water and Environment Journal* **35** (1), 335–348.

First received 7 April 2021; accepted in revised form 19 June 2021. Available online 2 July 2021