Environ Geochem Health (2024) 46:31
https://doi.org/10.1007/s10653-023-01778-3

ORIGINAL PAPER

Trihalomethane prediction model for water supply system based on machine learning and Log-linear regression

Hui Li · Yangyang Chu · Yanping Zhu · Xiaomeng Han · Shihu Shu

Received: 10 September 2023 / Accepted: 9 November 2023 / Published online: 16 January 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract Laboratory determination of trihalomethanes (THMs) is a very time-consuming task. Therefore, establishing a THMs model using easily obtainable water quality parameters would be very helpful. This study explored the modeling methods of the random forest regression (RFR) model, support vector regression (SVR) model, and Log-linear regression model to predict the concentration of total-trihalomethanes (T-THMs), bromodichloromethane (BDCM), and dibromochloromethane (DBCM), using nine water quality parameters as input variables. The models were developed and tested using a dataset of 175 samples collected from a water treatment plant. The results showed that the RFR model, with the optimal parameter combination, outperformed the Log-linear regression model in predicting the concentration of T-THMs ($N_{25}=82$–88%, $r_p=0.70$–0.80), while the SVR model performed slightly better than the RFR model in predicting the concentration of BDCM ($N_{25}=85$–98%, $r_p=0.70$–0.97). The RFR model exhibited superior performance compared to the other two models in predicting the concentration of T-THMs and DBCM. The study concludes that the RFR model is superior overall to the SVR model and Log-linear regression models and could be used to monitor THMs concentration in water supply systems.

Keywords Disinfection by-products · Predictive models · Machine learning algorithms · Random forest · Support vector regression

Introduction

Disinfection is a conventional method to ensure water safety during the water treatment process. Chlorine disinfection is widely used in drinking water systems due to its low cost and simple operation. However, chlorination disinfection has the disadvantage of reacting with certain organic compounds in water to produce disinfection by-products (DBPs) (Chen et al., 2023; Yang et al., 2021). Studies have shown that DBPs have cytotoxicity and reproductive toxicity, leading to abnormal embryonic development (Egwari et al., 2020; Kar & Senthilkumaran, 2020; Srivastav & Kaur, 2020; Zhang et al., 2023). Currently, hundreds of kinds of DBPs have been discovered, among which trihalomethanes (THMs) and haloacetic acids (HAAs) are the most concerned and commonly monitored (Dubey et al., 2020; Ozgur & Kaplan-Bekaroglu, 2022). Worldwide implementation of relevant policies and water quality standards strictly

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10653-023-01778-3.

H. Li · Y. Chu · Y. Zhu · X. Han · S. Shu (✉)
College of Environmental Science and Engineering, Donghua University, No. 2999 North Renmin Road, Shanghai 201620, China
e-mail: 2222377@mail.dhu.edu.cn

controls the concentration of DBPs. For example, the US Environmental Protection Agency has set a maximum level of 80 µg/L for total trihalomethanes (T-THMs), bromodichloromethane (BDCM), dibromochloromethane (DBCM), and tribromomethane (TBM). China also specifies that the sum of the measured concentrations of THMs and their respective limits should not exceed 1 ($CHCl_3$: 60 µg/L, DBCM: 100 µg/L, BDCM: 60 µg/L, TBM: 100 µg/L).

Although THMs testing is critical, the process is usually cumbersome and time-consuming. It requires not only expensive instrument equipment and experimental reagents but also complex pre-processing work (Liu et al., 2022; Mohammadi et al., 2020; Pérez-Lucas et al., 2022; Shi et al., 2023).

It is compressively known that the generation of THMs is closely related to disinfection conditions, including pH, disinfection time, chlorine dosage, and water quality parameters such as UV_{254}, dissolved organic carbon (DOC), temperature, Br^{-} (Hong et al., 2020; Liang & Singer, 2003). Therefore, many researchers used the relatively easy-to-obtain water quality parameters and disinfection conditions to establish a predictive model of DBPs (Albanakis et al., 2021; Peng et al., 2020; Uyak et al., 2005) in order to more easily monitor the concentration of DBPs. Multiple linear/nonlinear regression models are frequently used to predict THM concentrations. A predictive model for trihalomethanes was established using influencing factors, and the results showed that the stepwise regression model was superior to the least squares regression model and the multiple linear regression model (Albanakis et al., 2021). Another multiple regression model was established to predict the generation of THMs and haloacetonitriles (HANs) during the chlorination process of low SUVA source water, but the predictive performance of such models was unsatisfactory (Hong et al., 2016). Multiple regression models were built for THMs, HANs, and haloacetic acids (HAAs) in source water. The results showed that some models had high predictive accuracy, while others had low predictive accuracy (Lin et al., 2018). Uyak et al. (2005) established a multiple regression model to predict the THM concentration in the effluent, in which the result was accurate but the formula was relatively complex.

Recently, machine learning (ML) has gradually become a research hotspot in many fields, such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Random Forest (RF). However, there is little research on predicting THMs using ML methods. The research on DBPs prediction by different ML methods is summarized in Table 1. A small amount of research has used ANN to predict THMs, such as backpropagation neural networks (BPNN) and radial basis function neural networks (RBF ANN) (Hong et al., 2020; Liu et al., 2023). Compared to BPNN, RBF ANN performs better in predicting THM concentration. Hong et al. (2020) predicted THMs by combining RBF ANN with gray relational analysis (GRA). Compared with linear regression models, GRA can establish a well-performing RBF ANN model with fewer factors. Liu et al. (2023) compared the prediction effects of BPNN, genetic algorithm backpropagation neural network (GA-BPNN), and generalized regression neural network (GRNN) models on THMs. The study showed that BPNN had average prediction performance, GA-BPNN had good performance but a long prediction time, and GRNN had the best prediction performance. SVM is a supervised ML algorithm, while Support Vector Regression (SVR) is a data regression algorithm based on SVM. SVR can learn the relationship between complex data and solve nonlinear problems and is therefore increasingly applied in the field of water quality monitoring (Zheng et al., 2013). A large amount of research has found that the SVR model performs better than traditional multiple linear regression models. Meanwhile, RF is an ensemble algorithm (Ma et al., 2023), which can calculate simultaneously, greatly shorten the running time, and effectively process noisy data and outlier values. Based on the RF algorithm, Peng et al. (2023) compared the effects of support vector machine, random forest, and stepwise multiple linear regression on the prediction of emerging DBPs. The results showed that ML methods were more suitable for managing the generation of DBPs than stepwise multiple linear regression. Hu et al. (2023) studied the predictive performance of eleven machine learning models for emerging DBPs, and results showed that RF was the best model among the regression tree categories. However, RF is rarely used to predict DBPs, and it can determine the degree of importance of variables compared to SVM and ANN. Therefore, RF was further applied to the prediction of DBPs in this study. Moreover, the building process and parameter optimization process of most ML models are fuzzy, which makes it difficult for

Table 1 Summary of the application of ML methods in DBPs

ML methods	Objects	Input variable	Model performance	Advantages	Disadvantages	References
RBF ANN, GRA	T-THMs, TCM, BDCM	pH, temperature, chlorine, DOC, UVA_{254}, NH_4^+–N, NO_2^-–N, Br^-	0.760–0.925[a], 92–98% (RBF ANN) 0.760–0.946[a], 92–98%[b] (GRA)	High prediction accuracy and more detailed parameter optimization process. GRA can be a viable approach to develop resilient models with reduced variables	the number of data samples was small	Hong et al., 2020
SVR, RF	21 DBPs	Chlorine, UV_{254}, pH, TOC, Br^-	2.578–5.798%[d], 0.665–0.802[c] (SVR) 2.867–5.346%[d], 0.671–0.965[c] (RF)	The prediction model of chloral hydrate (CH) was developed, better prediction effect was obtained by using fewer water quality parameters	The standardization of generation conditions limited the application of the model and lacked aromatic DBPs	Peng et al., 2023
BPNN, GABP, GRNN	T-THMs, DBCM, BDCM	Temperature, pH, UV_{254}, TOC, COD, NH_4^+–N, NO_3^-–N, NO_2^-–N, Br^-	0.571–0.857[a], 61.5–91.5% [b] (BPNN) 0.573–0.863[a], 68.2–93.6%[b] (GABP) 0.657–0.824[a], 81.8–100%[b] (GRNN)	The three methods were compared systematically, a new method GRNN was used to predict DBPs, and the prediction effect was excellent	The temperature data was not the real water temperature but the weather data, and the input data lacked the residual chlorine	Liu et al., 2023
ANN, SVM, GEP	THMs	Temperature, pH, Br^-, Cl_2/DOC, contact time	0.918[c], 4.05[e](ANN) 0.935[c], 3.66[e](SVM) 0.933[c], 3.63[e](GEP)	The prediction performance of the model was excellent, SVM performed slightly better than ANN and GEP	The number of data samples was small	Singh et al. 2012
RSRM, ANFIS	TCM, BDCM, DBCM, T-THMs	Temperature, pH, UV_{254}, residual chlorine, DOC	0.727–0.886[c], 0.017–1.449[e], 0.01–0.978[f] (RSRM) 0.956–0.989[c], 0.007–0.401[e], 0.002–0.133[f] (ANFIS)	The prediction performance of ANFIS model was better than that of RBF ANN, the interaction between different water quality parameters and DBPs was analyzed	The sample size is too small, only 24	Okoji et al., 2022

Table 1 (continued)

ML methods	Objects	Input variable	Model performance	Advantages	Disadvantages	References
Eleven ML methods	THMs, HAAs, DCAN, CPK, TCP	Temperature, pH, UV_{254}, DOC, chlorine, turbidity, conductivity	$0.94^{c}$for DCAN, TCP(SVR) $0.92^{c}$for CPK(GPR)	Potential models were identified among eleven machine learning models	Lack of emerging disinfection byproducts	Hu et al., 2023
K-PLS, SVM	THMs	UV_{254}, residual chlorine, TOC, turbidity,	–	The prediction effect of K-PLS and SVM was better than MLR	–	Platikanov et al., 2012
ANN	THMs, HAAs, TOX	Temperature, pH, UV_{254}, TOC, chlorine, Br^{-}, contact time	0.78–0.98^{c}	ANN was used to predict DBPs after different treatments (conventional treatment, chlorination, granular activated carbon, and nanofiltration), the prediction effect was good	Lack of emerging disinfection byproducts	Kulkarni et al. 2010

a: Pearson correlation coefficient (r_p), b: The accuracy of the predicted values (N_{25}), which is the percentage of the absolute error $E < 25\%$, c: Coefficient of determination (R^2), d: Mean absolute percentage error (MAPE), e: Root mean square error (RMSE), f: Mean absolute error (MAE), ANFIS: Advanced neuro-fuzzy inference system, RSEM: Response surface regression model, GPR: Gaussian process regression

the models to be further applied in other conditions. These processes are explained in detail in this study, which is conducive to the further popularization and application of the model. Additionally, although some studies (Okoji et al., 2022; Singh et al., 2012) have achieved good predictive performance, the sample size of the dataset is small, which makes it unable to fully represent the overall sample distribution, thus limiting the reliability and generalization ability of the results. Therefore, this study expanded the sample size of the dataset, making the prediction results more convincing.

To further investigate the performance of ML on THMs, this study utilized water quality parameters collected from a water plant in Taihu Lake to establish Log-linear regression models, SVR models, and RFR models for predicting the concentration of T-THMs, DBCM, and BDCM. The models were then compared with actual measurements using evaluation indicators to assess their predictive effectiveness.

Materials and methods

Datasets and grouping

The data used in this study, including 175 samples from Water Plant B between 2016 and 2021 were all from Liu et al. (2023). Water quality data were collected from the inlet pipe of Water Plant B. THMs levels were sampled from the pipeline connected to the clean water reservoir of Plant B. The concentration ranges of THMs and corresponding water quality parameters are shown in Table 2. The water plant sources from Taihu Lake are treated by traditional processes including coagulation, sedimentation, filtration, activated carbon treatment with ozone, and chlorination. The data set includes T-THMs, DBCM, BDCM concentrations, and nine water quality parameters (temperature, pH, UV_{254}, TOC, COD, NH_4^+–N, NO_3^-–N, NO_2^-–N, Br^-). The USEPA 551.1 method was used to measure THMs, and the water quality parameters were determined by the standard method (APHA, 1998).

THMs generation is related to multiple water quality parameters. Temperature reflects the seasonal changes in water temperature, and pH affects the removal of organic matter, thereby affecting the generation of THMs. Organic matter is an important

Table 2 Concentration range of THMs and corresponding water quality index of waterworks B

Values	Units	Data set	
		Training set	Test set
T-THMs	μg /L	13–70	21–62
DBCM	μg /L	3.2–23.9	6.8–23
BDCM	μg /L	2.6–15.4	4.6–12.3
pH	/	7.3–8.2	7.4–8
COD	mg/L	2.48–6.2	2.99–5.75
NH_4^+-N	mg/L	0.05–0.43	0.06–0.24
NO_2^--N	mg/L	0.003–0.105	0.006–0.089
NO_3^--N	mg/L	0.1–2.32	0.12–0.46
Br^-	mg/L	0.05–0.197	0.065–0.177
TOC	mg/L	2.875–6.416	2.93–5.811
UV_{254}	A/cm	0.059–0.314	0.065–0.246
Temp	℃	4–37	9–36

precursor to the formation of THMs, and its concentration is often characterized by COD, TOC, and UVA_{254}, NH_4^+–N, NO_3^-–N, NO_2^-–N reduce the concentration of THMs by consuming free chlorine in water, while the presence of Br^- increases the concentration of brominated THMs. Therefore, nine water quality parameters were selected as input variables to predict THMs concentrations.

The data set was divided into a training set and a test set. In this respect, 80% (141 samples) were used as the training set and 20% (34 samples) as the test set. The maximum and minimum samples of T-THMs, DBCM, and BDCM were added to the training set to ensure the predictive ability of the model, while the remaining data were randomly assigned to the training and test sets.

Establishment of Log-linear regression model

The multiple nonlinear regression model for THMs was established using SPSS 26.0. The concentration ($\log_{10} Y_i$) of T-THMs, DCBM, and BDCM from 141 training was set as dependent variables (Y), and the logarithmic values ($\log_{10} X_i$) of nine water quality parameters as independent variables (X). The equation was established using a stepwise regression method. For example, the equation for T-THMs was $\log_{10}$T-THMs $= \log_{10} k_0 + k_1 \log_{10} X_1 + k_2 \log_{10} X_2 + \ldots + k_i \log_{10} X_i$, ($i = 1, 2,, 9$). Finally, the equation was transformed into T-THMs $= 10^{k_0} X_1^{k_1} \ldots X_i^{k_i}$, where $k_0, k_1, \ldots, k_i$ are constants.

Establishment of SVR model

SVM is a binary classification algorithm based on ML theory that can effectively find the globally optimal solution (Vapnik, 1995; Zheng et al., 2023). In Fig. 1, the basic principle is to map the sample data through a non-linear kernel function to a high-dimensional space, find a few effective support vectors to represent the entire dataset, and make the data linearly separable in the high-dimensional space. By learning from a limited amount of data, the optimal solution can be obtained (Q. Xu et al., 2015).

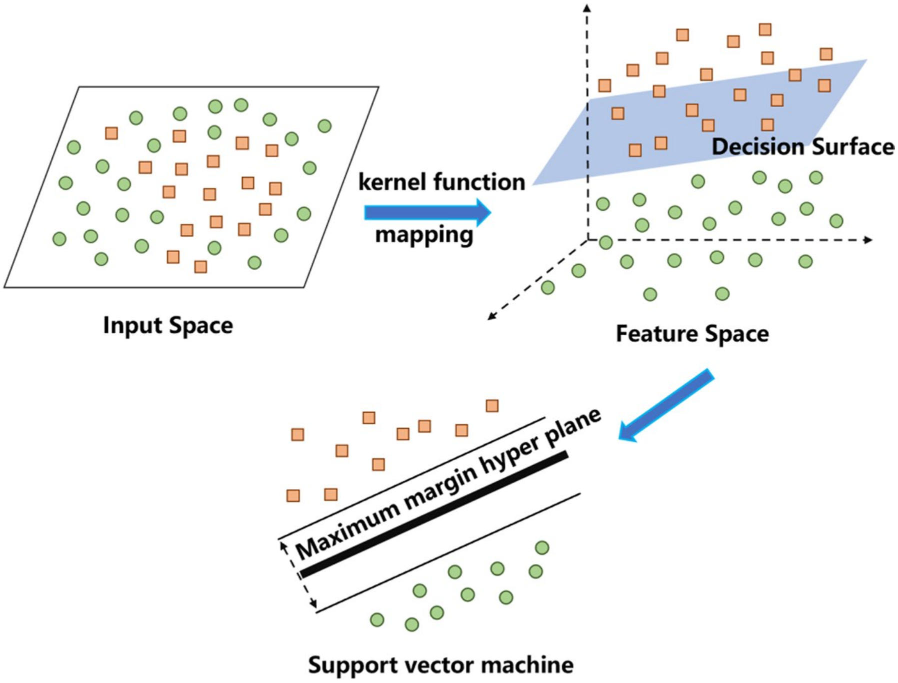
SVR differs from SVM because SVM primarily finds an optimal hyperplane that separates two or more classes of samples, commonly used for data classification. On the other hand, SVR requires that the total deviation of all sample points from the optimal hyperplane is minimized (Peng & Chen, 2018). For a given dataset (x_i, y_i)(with x_i as input variables, y_i as corresponding output variables, $i = 1, 2, \ldots, l$), the optimal linear regression function is constructed in a high-dimensional space:

$$f(x) = \omega^T \varphi(x) + b \tag{1}$$

where $\varphi(x)$, ω, b represents nonlinear mapping kernel function, weight vector and threshold value, respectively.

SVR introduces the insensitive loss function ε based on SVM, representing the boundary error of the decision function in high-dimensional space. The prediction accuracy of the SVR model mainly depends on the penalty factor c, kernel parameter g, and ε. c represents the punishment intensity for sample errors exceeding ε, where a large c indicates an emphasis on errors, but may result in overfitting, while a small c indicates a high tolerance for errors but may lead to underfitting. g is used to adjust the shape and range of the radial basis function, affecting the decision boundary of the model. ε affects the accuracy of the regression model (Abu Awad et al., 2017), and the larger the value, the lower the error tolerance, resulting in relatively high accuracy. An optimal training model with specific values of c and g was obtained through grid search and manual tuning. The search range for c was set to [0.1, 2] with a step size of 0.1,

Fig. 1 Support vector machine (SVM) schematic diagram



while the search range for g was [0.1, 8] with a step size of 0.1. The R^2 and RMSE of the training set and test set of the models under different parameter combinations were calculated and then output into Excel tables. For further comparison, 5 groups of parameter combinations whose R^2 is closest to 1 and RMSE is smaller were selected, and then the optimal parameter combinations of the model were found by manual tuning.

The SVR model was constructed using MATLAB R2018a, which mainly includes seven steps in the modeling process, Fig. 2 shows the main steps. The "L2 regularization" in the third step is a mathematical technique employed to mitigate model overfitting and enhance generalization capability. It is frequently employed to fine-tune the trade-off between the error term and regularization coefficient in SVR.

Establishment of RFR model

The RF algorithm is an ensemble learning method that builds multiple decision trees related to random vectors to classify or regress samples. It is capable of handling outliers and noise effectively. In Fig. 3, the basic principle of the RF algorithm is to generate multiple decision trees randomly by sampling from the training set. When making regression predictions, all decision trees predict each sample, and the average is taken as the predicted result of the sample (Karabadji et al., 2023).

The steps of the RF algorithm are as follows: (1) Assume a training set $T = \{(x_1, y_1), ..., (x_n, x_n)\}$, then generate a random sequence $\delta_i (i = 1, ..., k)$, and re-sample k new sample sets from the training set T through bootstrapping. (2) Generate a decision tree $[g(X, \delta_i), i = 1, ..., k]$ for each new sample set, where matrix X is the independent variable of the model. (3) Generate all decision trees completely and obtain the random forest model sequence $[g_1(X), g_2(X), ..., g_k(X)]$. (4) Input the new sample into the decision tree, obtain the predicted result of each decision tree, and take the average value to obtain the predicted value of the random forest:

$$G(x) = 1/k \sum_{i=1}^{k} g_i(x) \tag{2}$$

where $G(x)$, $g_i(x)$, k are the final prediction result, the prediction results for a single decision tree, and the number of decision trees, respectively.

The RFR model was established using MATLAB R2018a. Nine water quality parameters, similar to those used in the SVR model, were selected as input variables for the model, while three types of THMs were used as output variables. Based on the RF algorithm

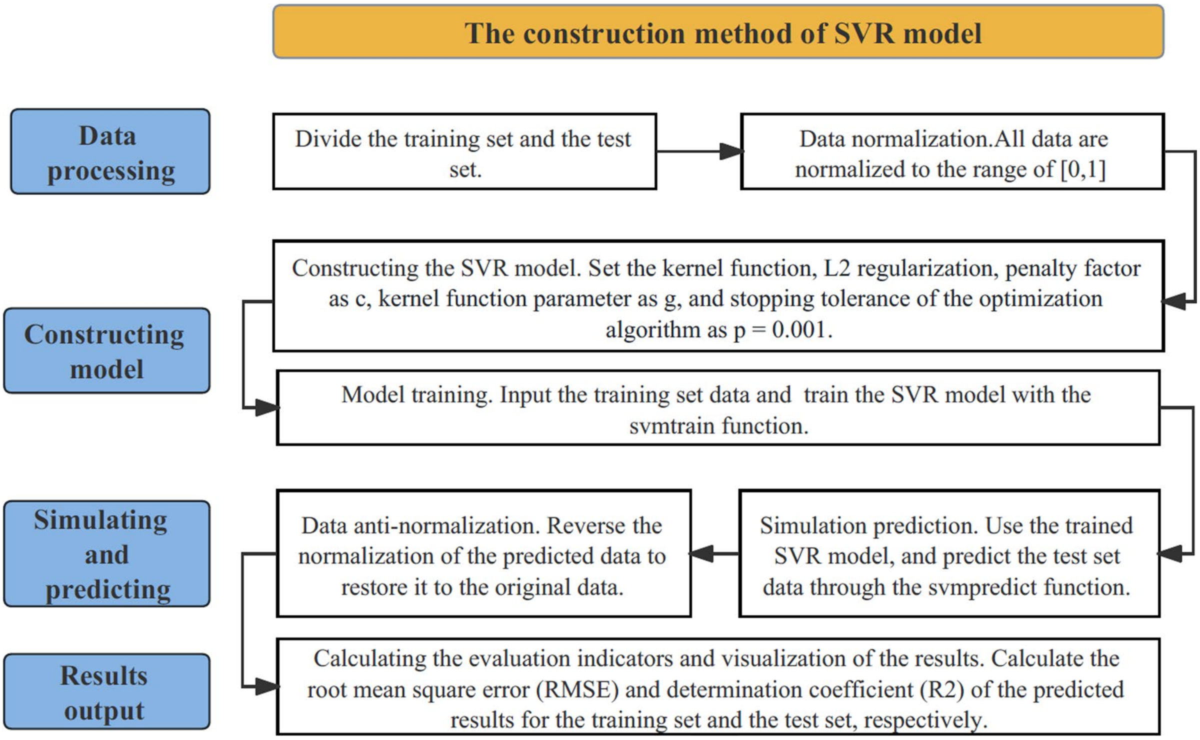
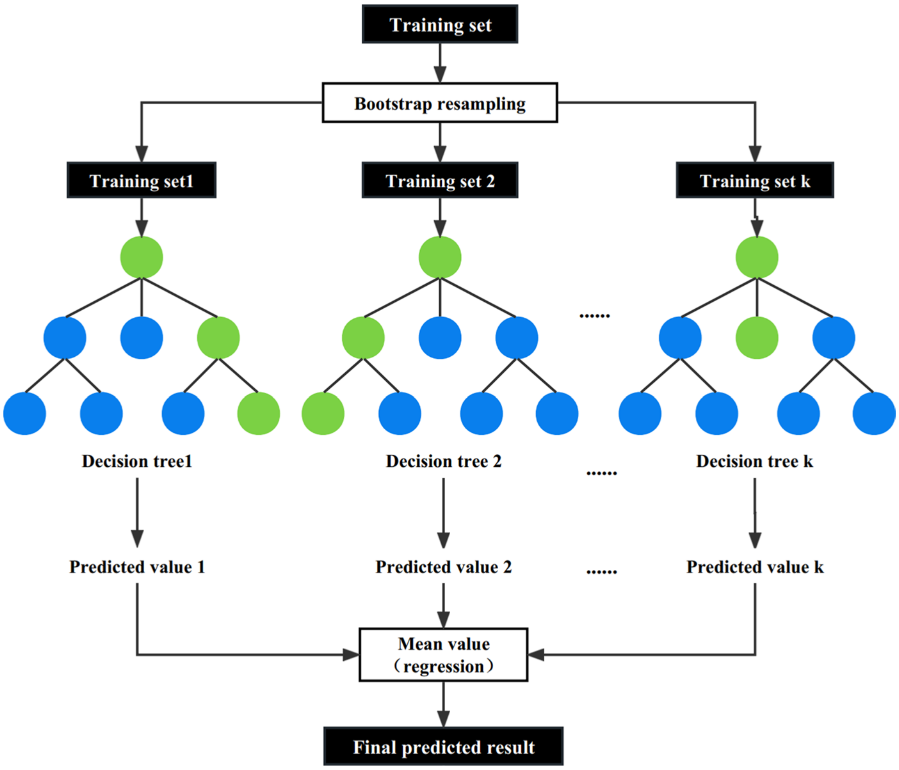

The construction method of SVR model
Data processing
Divide the training set and the test set.
Data normalization.All data are normalized to the range of [0,1]
Constructing model
Constructing the SVR model. Set the kernel function, L2 regularization, penalty factor as c, kernel function parameter as g, and stopping tolerance of the optimization algorithm as p = 0.001.
Model training. Input the training set data and train the SVR model with the svmtrain function.
Simulating and predicting
Data anti-normalization. Reverse the normalization of the predicted data to restore it to the original data.
Simulation prediction. Use the trained SVR model, and predict the test set data through the svmpredict function.
Results output
Calculating the evaluation indicators and visualization of the results. Calculate the root mean square error (RMSE) and determination coefficient (R2) of the predicted results for the training set and the test set, respectively.

Fig. 2 SVR model construction method

Fig. 3 Random forest regression (RFR) schematic diagram



principle, the TreeBagger function was used to train the RFR model, setting parameters such as the number of decision trees M and the minimum leaf number N. A grid search method was used to determine values for *M* and *N*, the range for *M* was [10,60] with a step size of 1, and *N* was [1,10] with a step size of 0.1. The subsequent steps were the same as those for the SVR model.

Model evaluation

To evaluate the error between the predicted values and the measured values of the three models, the mean absolute error (MAE) and Pearson correlation coefficient (r_p) were used as the model evaluation indicators. Following the research conducted by Xu et al. (2022), the absolute error (E) between the predicted values and the measured values was calculated, and the accuracy of the predicted values was defined as N_{25}, which is the percentage of $E < 25\%$. The low value of MAE and the high values of N_{25} and r_p represent the high accuracy of the model.

$$E = abs(\{[THMs]_{predicted} - [THMs]_{measured}\} / [THMs]_{measured}) \times 100\% \tag{3}$$

$$MAE = 1/m \sum_{i=1}^{m} (|y_i - f(x_i)|) \tag{4}$$

where y_i, $f(x_i)$, and m represents the measured values, the predicted values, and the number of samples, respectively.

Results and discussion

Log-linear regression model

Analysis of Log-linear regression model

Regression models of various THMs generated by training sets are shown in Table 3.

The F-test results of all models (S-Table 1) exceeded the F-test critical values at 99% confidence levels ($F_{\alpha=0.01}(3137) = 3.928$, $F_{\alpha=0.01}(4136) = 3.460$), indicating a significant linear relationship between THMs and water quality parameters. Moreover, all models had variance inflation factors (VIF) much less than 5, indicating the absence of multicollinearity problems among the variables.

Models (1) and (2) had three important water quality parameters, while model (3) had four. In this study, T-THMs, DBCM, and BDCM showed a positive correlation with temperature, as increasing temperature will accelerate the reaction rate and thus affect THM generation. T-THMs, DBCM, and BDCM were negatively correlated with NO_3^-–N, indicating that NO_3^-–N affected THMs concentration by consuming available chlorine. T-THMs and DBCM were positively correlated with Br^- while Br^- had a greater impact on DBCM than T-THMs. Only the regression model for BDCM included UV_{254}, indicating that UV_{254} has a certain influence on BDCM formation, which was consistent with the results of this research (Xu et al., 2022). Furthermore, model (3) also included COD, whose impact on BDCM formation was lower than temperature.

Evaluation of log-linear regression model

The testing set, training set, and complete set were input into the models (1) ~ (3), and compared with the measured values. Correlation analysis was conducted in Fig. 4.

In Fig. 4(a1), (b1) and (c1), it can be inferred that for the complete set, the N_{25} values of T-THMs, DBCM, and BDCM were 77–81%, while the r_p values were 0.65–0.78. In Fig. 4(a2), (b2) and (c2), regarding the training set, the N_{25} values of T-THMs, DBCM, and BDCM were 77–79%, and the r_p values were 0.66–0.79, which demonstrated comparable performance to the complete set. In Fig. 4(a3), (b3) and (c3), for the testing set, the N_{25} value of T-THMs reached 85%, while those of DBCM and BDCM were

Table 3 Log-linear regression models of various THMs

Number	Equations	R^2	F	P
(1)	$T - THMs = 10^{1.128}T^{0.442}(NO_3^- - N)^{-0.108}(Br^-)^{0.183}$	0.583	66.199	< 0.01
(2)	$DBCM = 10^{0.91}T^{0.436}(NO_3^- {-} N)^{-0.08}(Br^-)^{0.436}$	0.586	67.108	< 0.01
(3)	$BDCM = 10^{0.746}T^{0.466}(NO_3^- - N)^{-0.139}(UV_{254})^{0.327}DOC^{-0.34}$	0.496	35.486	< 0.01

both 76%. However, the r_p values ranging from 0.62 to 0.76 were not ideal, and the overall predictive performance was not satisfactory. The predictive ability of the Log-linear model was relatively intermediate.

Support vector regression model (SVR)

Determination of model parameters

Nine water quality parameters were selected as input variables, and T-THMs, DBCM, and BDCM were selected as output variables. Parameters obtained from grid search and corresponding tests-RMSE are shown in Fig. 5. The parameters with the smallest RMSE and the R^2 closest to 1 for both the training and testing sets were selected. When c was in the range of [0.3, 2] and g was in the range of [6, 8], the RMSE of T-THMs and DBCM in the testing set could be minimized. When c was in the range of [0.5, 1] and g was in the range of [8, 10], the RMSE of BDCM in the testing set could be minimized.

For T-THMs, when $c = 1.1$ and $g = 8$, the RMSE of the training set and the test set were both small, and R^2 was closest to 1. The same parameter optimization method was used for DBCM, and the value range of g for BDCM was adjusted to [0.1,

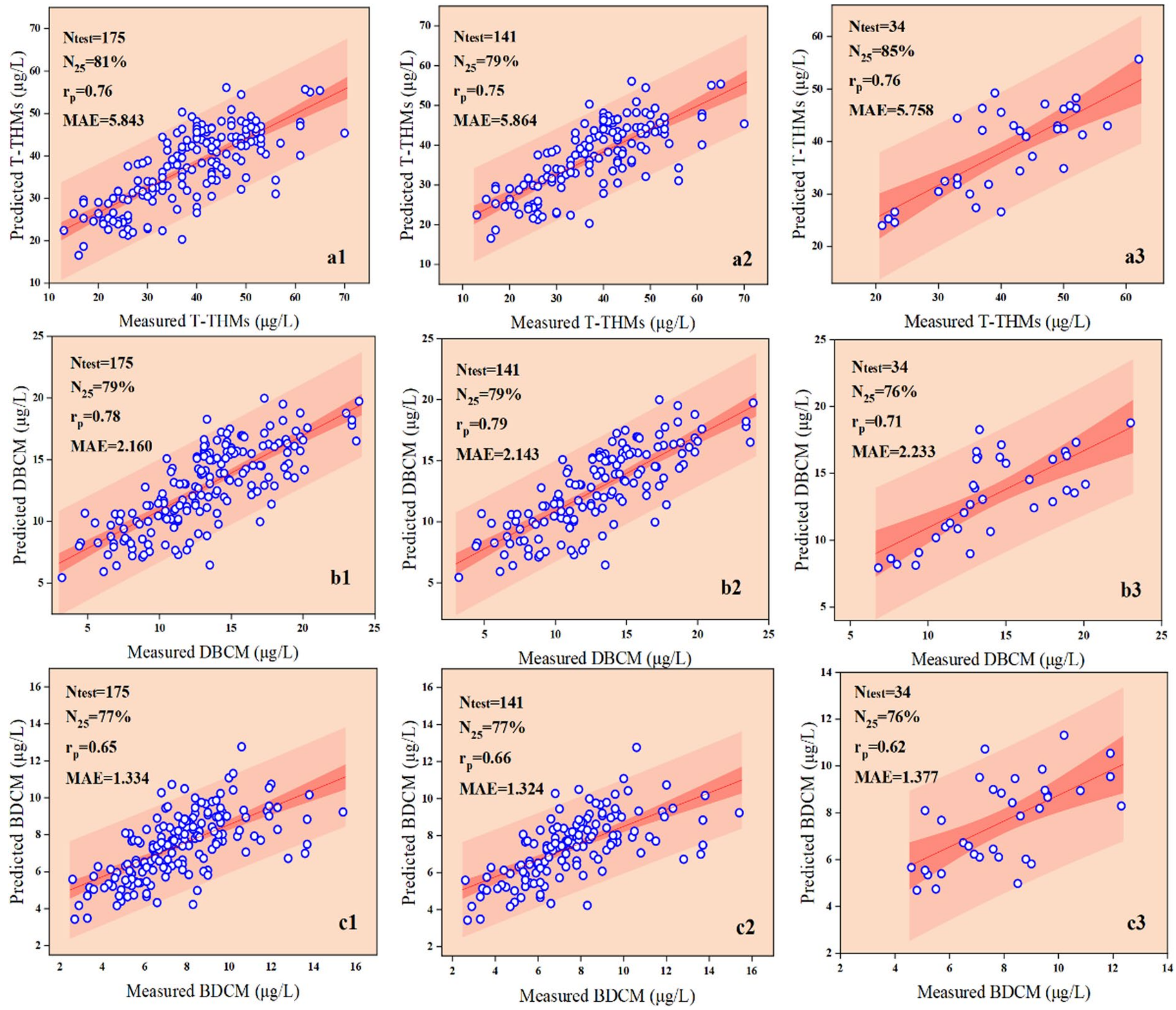
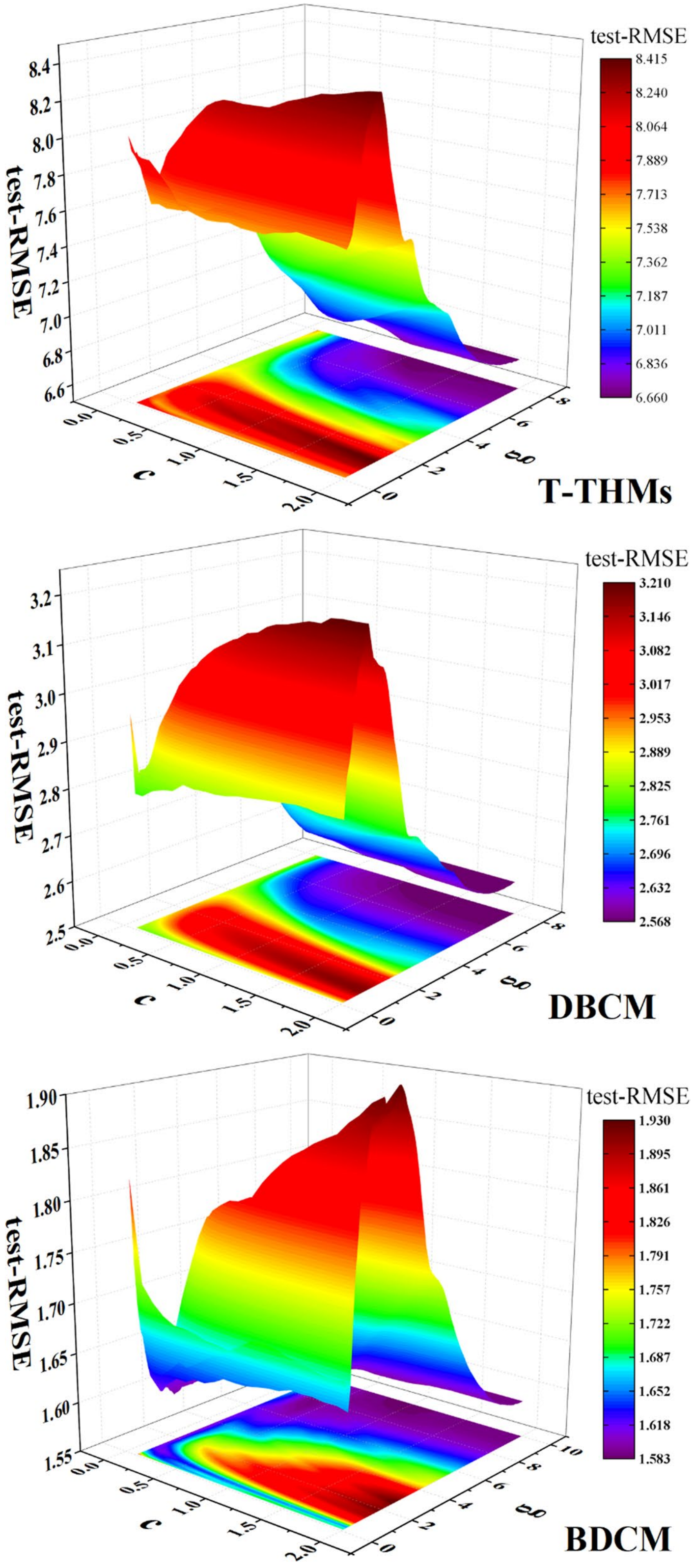


Fig. 4 Comparison of the predicted and measured values of the Log-linear regression model. a, b, and c represent T-THMs, DBCM, and BDCM. 1, 2, and 3 represent the complete set, training set, and test set, respectively

Fig. 5 RMSE under different c, g parameter combinations



10] with a step size of 0.1. In S-Tab.2, the best values of *c*, *g* and *p* for T-THMs, DBCM, and BDCM were obtained.

Prediction of THMs concentration

The fitting effect of the SVR model with optimal parameters on THMs is shown in Fig. 6(a1–c1), (a2–c2). The simulation effect of the complete set and training set was better than that of the Log-linear regression model. Under the complete set simulation, the N_{25} of T-THMs, DBCM, and BDCM were 94–96%, and the r_p of predicted values and measured values were 0.94, which were higher than the N_{25} (77–81%) and r_p (0.65–0.78) of the Log-linear regression model. Under the training set, the N_{25} of T-THMs, DBCM, and BDCM were 96–98%, and the r_p of predicted values and measured values were 0.95–0.97.

Following the utilization of the trained model to make predictions on the test set, the predicted performance of THMs is shown in Fig. 6(a3), (b3) and (c3). The SVR model showed improved performance in the test set ($N_{25}=82–85\%$), but its r_p (0.69–0.75) was comparable to that of the log-linear regression model. In addition, the SVR model exhibited overfitting, possibly due to the inadequate settings of the parameters *c* and *g*, which resulted in the insufficient generalization ability of the model.

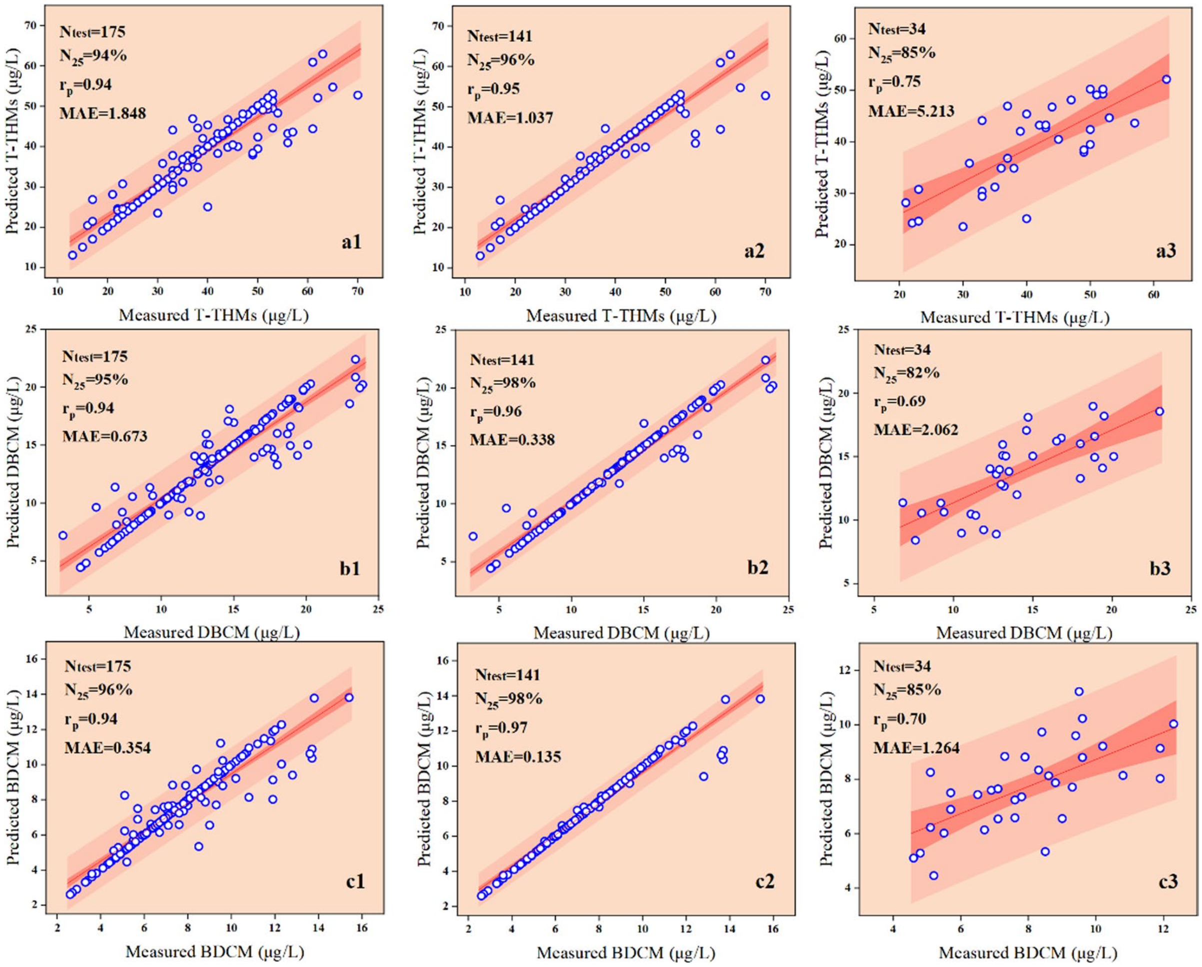


Fig. 6 Comparison of the predicted and measured values of the SVR model. a, b, and c represent T-THMs, DBCM, BDCM; 1, 2, and 3 represent the complete set, training set, and test set, respectively

Random forest regression model (RFR)

Determination of model parameters

The RFR model has two main adjustable parameters, namely the number of decision trees (M) and the minimum number of leaf nodes (N). Due to the stochastic nature of predictive performance in the RFR model, three sets of parameter combinations resulting in good predictive parameters were selected for further training. The optimal parameter combinations for THMs were obtained as shown in S-Table 3.

Prediction of THMs concentration

When using the optimal parameter combination, the RFR model exhibited a fitting effect on THMs, as shown in Fig. 7. In Fig. 7(a3), (b3) and (c3), the predicted results of the test set indicated a high regression coefficient r_p (0.70–0.80) and N_{25} (82–88%) between the predicted values and the measured values by the RFR model. In Fig. 7(a1), (b1) and (c1), for the complete set, $r_p = 0.92–0.95$ and $N_{25} = 91\%$-96%. In Fig. 7(a2), (b2) and (c2), for the training set, $r_p = 0.96–0.98$ and $N_{25} = 94–98\%$. Overall, the predictive ability of the RFR model for the prediction of THMs was better than the results of the Log-linear regression model, which obtained a r_p of 0.62–0.78 and a N_{25} of 76–85%.

Model comparison

Based on the evaluation indicators from Table 4, it can be seen that the prediction accuracy of the RFR model for the T-THMs, DBCM, and BDCM in the test set ($N_{25} = 82–88\%$, $r_p = 0.70–0.80$) was higher than that of the Log-linear regression model ($N_{25} = 76–85\%$, $r_p = 0.62–0.76$), indicating that the prediction ability of the RFR model was greater than the Log-linear regression model. In the prediction of the THMs of the complete set and training set, the prediction effect of the RFR model ($N_{25} = 91–98\%$, $r_p = 0.92–0.98$) was also better than the Log-linear regression model ($N_{25} = 77–81\%$, $r_p = 0.65–0.79$), which further indicated that the RFR model could better predict the concentration of THMs under the influence of multiple water quality factors. In the complete set, the MAE for T-THMs in the RFR model was 1.016, which was lower than that of 5.843 in the Log-linear regression model and 1.848 in the SVR model. However, for DBCM and BDCM, the MAE of the SVR model was lower than the other two models. The performance of SVR model for prediction of BDCM was better than the results of the RFR and Log-linear regression models which obtained MAEs of 0.702 and 1.334.

To better compare the performance of the Log-linear regression model, SVR model, and RFR model, the prediction results on the test sets of T-THMs, DBCM, and BDCM are shown in Fig. 8. All three models conformed to the overall trend of observation. Among them, the RFR model was more accurate in fitting T-THMs and DBCM compared to other models. For BDCM, the deviation of the three models from the measured values was relatively large, especially for noise points with a large deviation, and it was not enough to determine which model was superior based on the figure.

In Fig. 9, the box plot displayed the absolute relative errors between predicted and measured values for the three models on the test set. The values at the bottom of the figure represent the mean and median values of the absolute relative errors. A box-plot generally measures the dispersion of data using the interquartile range (IQR). It can be observed that BDCM had the most outliers among the three THMs, with three outliers, while T-THMs had one outlier. However, the number of outliers for all THMs was within a reasonable range. For BDCM, the RFR model had the smallest mean value of 0.141 and median value of 0.107. However, the IQR value of the RFR model was higher than the SVR model, indicating that predictions for BDCM in the RFR model were more dispersed than those of the SVR model. For DBCM, the RFR model had the smallest mean of 0.127 and a median of 0.09. The IQR value was also the lowest, indicating that the RFR model had the highest overall prediction quality compared to the other two models. For T-THMs, the RFR model had a smaller mean and lower IQR value than the SVR model, indicating higher stability of the prediction results. However, the median value was larger than that of the SVR model. Further analysis of Figs. 6 and 7 revealed that for T-THMs, the RFR model's N_{25} (88%) was greater than the SVR model's N_{25} (85%), and the r_p (0.79) was also greater than the SVR model's r_p (0.75). Therefore, for T-THMs, the RFR model's predictive performance was superior to the SVR model.

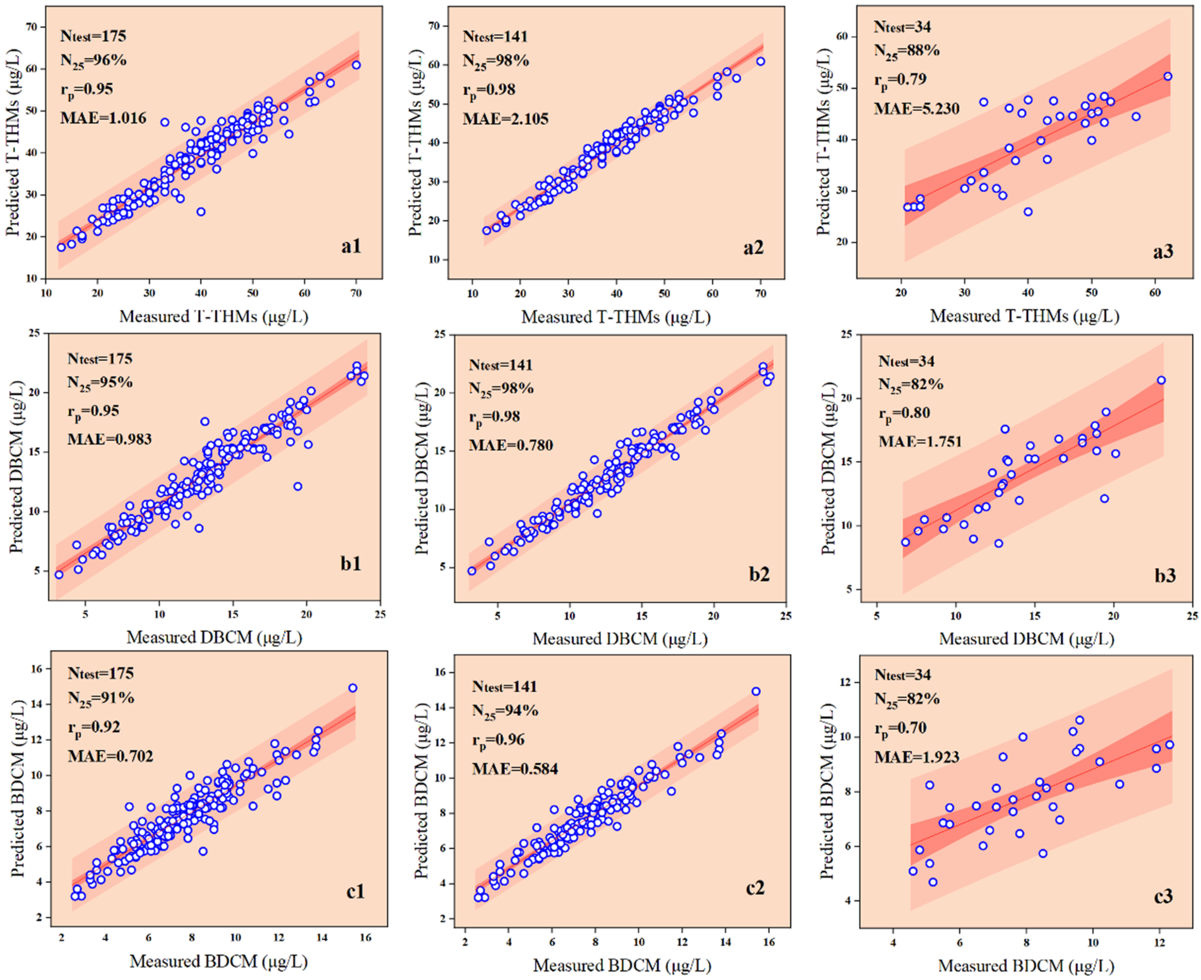


Fig. 7 Comparison of the predicted and measured values of the RFR model. a, b, and c represent T-THMs, DBCM, BDCM; 1, 2, and 3 represent the complete set, training set, and test set, respectively

Table 4 Evaluation indicators of three models under optimal parameter combination

THMs	Model	Complete set			Training set			Test set		
		N_{25}	r_p	MAE	N_{25}	r_p	MAE	N_{25}	r_p	MAE
T-THMs	LLR	81%	0.76	5.843	79%	0.75	5.864	85%	0.76	5.758
	SVR	94%	0.94	1.848	96%	0.95	1.037	85%	0.75	5.213
	RFR	96%	0.95	1.016	98%	0.98	2.105	88%	0.79	5.230
DBCM	LLR	79%	0.78	2.160	79%	0.79	2.143	76%	0.71	2.233
	SVR	95%	0.94	0.673	98%	0.96	0.338	82%	0.69	2.062
	RFR	95%	0.95	0.983	98%	0.98	0.780	82%	0.80	1.751
BDCM	LLR	77%	0.65	1.334	77%	0.66	1.324	76%	0.62	1.377
	SVR	96%	0.94	0.354	98%	0.97	0.135	85%	0.70	1.262
	RFR	91%	0.92	0.702	94%	0.96	0.584	82%	0.70	1.923

LLR represents the Log-linear regression model

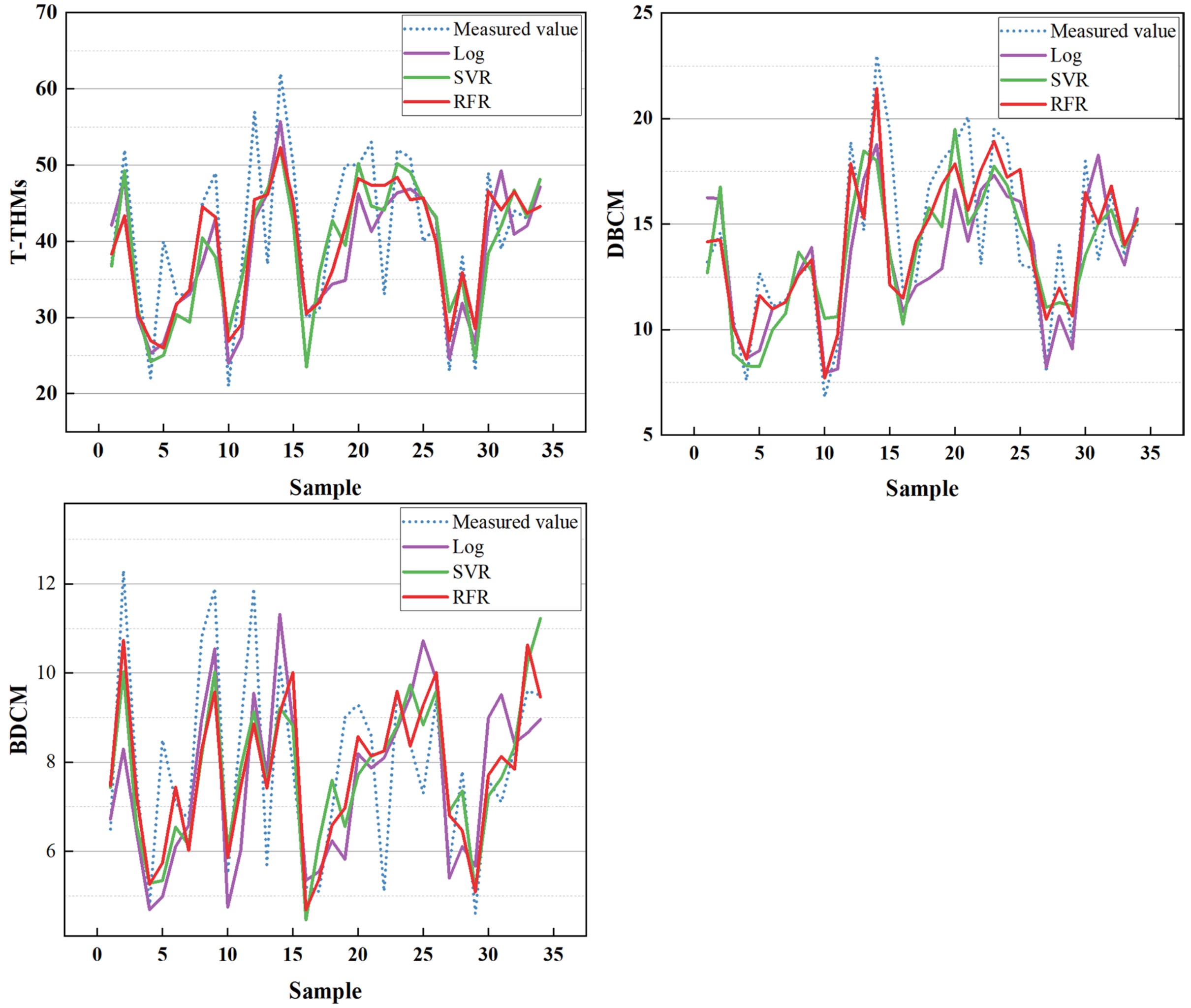
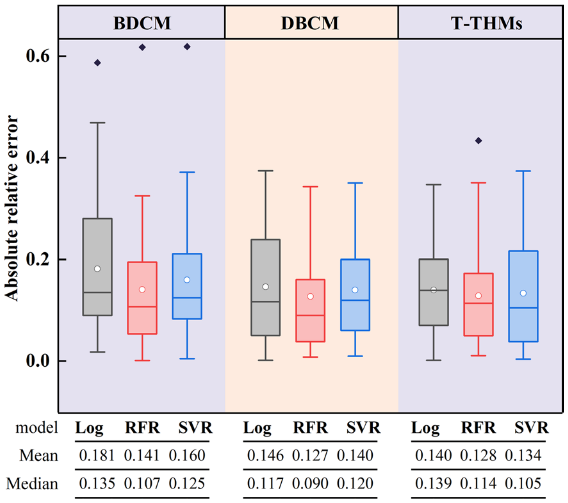


Fig. 8 The comparation of measured results and predicted results of the test sets of the three models

The predictive accuracy of models is generally positively correlated with the technical complexity of their construction (Hu et al., 2023). Although the accuracy of Log-linear regression models is not high, these models can provide specific formulas and good interpretability, which helps explain the relationship between different water quality parameters and DBPs. On the other hand, ML models such as BPNN, CNN, SVR, and RFR do not possess interpretability, but they are able to provide higher accuracy.

In this study, compared to the Log-linear regression model, the SVR and RFR models were relatively simple but had better predictive performance for three types of THMs. For T-THMs, the average absolute relative errors of the SVR and RFR models were reduced by 21% and 38%, respectively. For DBCM, the SVR and RFR models resulted in a decrease of 22% and 66%, respectively. For BDCM, the average absolute relative error of the SVR and RFR models decreased by 43% and 79%, respectively. Based on the above analysis, the RFR model showed better performance than the SVR and Log-linear regression models for DBCM and T-THMs. However, for the prediction of BDCM, the SVR and RFR models had their strengths. Overall, for the THMs studied in this research, the performance of the RFR model was better than the SVR and Log-linear regression models. In addition, the r_p values of the SVR, RFR, and Log-linear regression models were 0.70, 0.70, and 0.62, respectively, all of which were higher than the GANN



model	Log	RFR	SVR	Log	RFR	SVR	Log	RFR	SVR
Mean	0.181	0.141	0.160	0.146	0.127	0.140	0.140	0.128	0.134
Median	0.135	0.107	0.125	0.117	0.090	0.120	0.139	0.114	0.105

Fig. 9 The absolute relative error of the three models applied to the test set

model (r_p = 0.473, DWTPs B) (Liu et al., 2023). This shows that the SVR model and RFR model have excellent performance in predicting DBPs, and further indicates that the parameter optimization method used in this study is feasible.

Conclusion

This study utilized nine water quality parameters (temperature, pH, UV_{254}, TOC, COD, NH_4^+–N, NO_3^-–N, NO_2^-–N, Br^-) as input variables to establish the Log-linear regression model, SVR model, and RFR model to predict the concentrations of THMs (T-THMs, DBCM, and BDCM). By comparing the predicted values with the measured values, the performance of the three models was analyzed. From the comparison results of the model evaluation indicators (R^2, MAE, and RMSE), the results showed that the prediction accuracy of the Log-linear regression model was not ideal. However, the model can explain the relationship between water quality parameters and THMs, which helps to accurately take measures to control the generation of THMs. The SVR model achieved a good prediction effect for the training set, but the prediction effect for the test set was not ideal, while the RFR model had an excellent prediction effect for both the training set and the test set. Among them, the RFR model had a better predictive ability for T-THMs and DBCM than other models, with higher correlation coefficients and prediction accuracy, indicating that the RFR model could handle the complex relationship between T-THMs and DBCM concentration generation and water quality parameters. For BDCM, the prediction effect of the RFR model and the SVR model was similar, and the model performance had its strengths. Although the average error of the RFR model is smaller than that of the SVR model, the distribution difference of the prediction error in different parts is larger than that of the SVR model, which also means that the prediction accuracy and stability of the model may have potential risks. Overall, the prediction effect of the RFR model was better than the SVR model.

Acknowledgements Financial support was received from the Key Program of the Shanghai Science and Technology Commission (19DZ1204401).

Author contribution H.L.: Acquisition and analysis of data; Methodology; Model testing; Writing. Y.C.: Revising it critically for important intellectual content. Y.Z.: Methodology; Polish; Final approval of the version to be submitted. X.H.: Polish. S.S.: Investigation; Polish.

Funding This work was supported by the Key Program of the Shanghai Science and Technology Commission (19DZ1204401).

Data availability Data will be made available on request.

Declarations

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

References

Abu Awad, Y., Koutrakis, P., Coull, B. A., & Schwartz, J. (2017). A spatio-temporal prediction model based on support vector machine regression: Ambient Black Carbon in three New England States. *Environmental Research, 159*, 427–434. https://doi.org/10.1016/j.envres.2017.08.039

Albanakis, C., Tsanana, E., & Fragkaki, A. G. (2021). Modeling and prediction of trihalomethanes in the drinking water treatment plant of Thessaloniki, Greece. *Journal of Water Process Engineering, 43*, 102252. https://doi.org/10.1016/j.jwpe.2021.102252

Chen, H., Lin, T., Wang, P., Zhang, X., Jiang, F., & Wang, Y. (2023). Novel solar/sulfite advanced oxidation process for carbamazepine degradation: Radical chemistry, transformation pathways, influence on disinfection byproducts and toxic changes. *Chemical Engineering Journal, 451*, 138634. https://doi.org/10.1016/j.cej.2022.138634

Dubey, S., Gusain, D., Sharma, Y. C., Bux, F. (2020). Chapter 15 - The occurrence of various types of disinfectant by-products (trihalomethanes, haloacetic acids, haloacetonitrile) in drinking water. In M. N. V. Prasad (Ed.), Disinfection By-products in Drinking Water (pp. 371–391). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-08-102977-0.00016-0

Egwari, L. O., Benson, N. U., Effiok, W. W. (2020). Chapter 8 - Disinfection by-product-induced diseases and human health risk. In M. N. V. Prasad (Ed.), Disinfection By-products in Drinking Water (pp. 185–204). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-08-102977-0.00008-1

Hong, H., Song, Q., Mazumder, A., Luo, Q., Chen, J., Lin, H., & Liang, Y. (2016). Using regression models to evaluate the formation of trihalomethanes and haloacetonitriles via chlorination of source water with low SUVA values in the Yangtze River Delta region, China. *Environmental Geochemistry and Health, 38*(6), 1303–1312. https://doi.org/10.1007/s10653-016-9797-1

Hong, H., Zhang, Z., Guo, A., Shen, L., Sun, H., Liang, Y., & Lin, H. (2020). Radial basis function artificial neural network (RBF ANN) as well as the hybrid method of RBF ANN and grey relational analysis able to well predict trihalomethanes levels in tap water. *Journal of Hydrology, 591*, 125574. https://doi.org/10.1016/j.jhydrol.2020.125574

Hu, G., Mian, H. R., Mohammadiun, S., Rodriguez, M. J., Hewage, K., & Sadiq, R. (2023). Appraisal of machine learning techniques for predicting emerging disinfection byproducts in small water distribution networks. *Journal of Hazardous Materials, 446*, 130633. https://doi.org/10.1016/j.jhazmat.2022.130633

Hydrophobic organic compounds in drinking water reservoirs: Toxic effects of chlorination and protective effects of dietary antioxidants against disinfection by-products. (2019). Water Research, 166, 115041. https://doi.org/10.1016/j.watres.2019.115041

Kar, S., Senthilkumaran, B. (2020). Chapter 16—Water disinfection by-products cause acute toxicity in teleosts: a review. In M. N. V. Prasad (Ed.), Disinfection By-products in Drinking Water (pp. 393–411). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-08-102977-0.00017-2

Karabadji, N. E. I., Amara Korba, A., Assi, A., Seridi, H., Aridhi, S., & Dhifli, W. (2023). Accuracy and diversity-aware multi-objective approach for random forest construction. *Expert Systems with Applications, 225*, 120138. https://doi.org/10.1016/j.eswa.2023.120138

Kulkarni, P., & Chellam, S. (2010). Disinfection by-product formation following chlorination of drinking water: Artificial neural network models and changes in speciation with treatment. *Science of the Total Environment, 408*(19), 4202–4210. https://doi.org/10.1016/j.scitotenv.2010.05.040

Liang, L., & Singer, P. C. (2003). Factors influencing the formation and relative distribution of haloacetic acids and trihalomethanes in drinking water. *Environmental Science and Technology, 37*(13), 2920–2928. https://doi.org/10.1021/es026230q

Lin, J., Chen, X., Ansheng, Z., Hong, H., Liang, Y., Sun, H., & Chen, J. (2018). Regression models evaluating THMs, HAAs and HANs formation upon chloramination of source water collected from Yangtze River Delta Region, China. *Ecotoxicology and Environmental Safety, 160*, 249–256. https://doi.org/10.1016/j.ecoenv.2018.05.038

Liu, B., Zheng, X., Ke, Y., Cao, X., Sun, Q., & Wu, H. (2022). Automated headspace solid-phase microextraction-gas chromatography-mass spectrometry of trihalomethane and typical nitrogenous disinfection by-products in water. *Journal of Chromatography A, 1673*, 463068. https://doi.org/10.1016/j.chroma.2022.463068

Liu, K., Lin, T., Zhong, T., Ge, X., Jiang, F., & Zhang, X. (2023). New methods based on a genetic algorithm back propagation (GABP) neural network and general regression neural network (GRNN) for predicting the occurrence of trihalomethanes in tap water. *Science of the Total Environment, 870*, 161976. https://doi.org/10.1016/j.scitotenv.2023.161976

Ma, X., Chen, Z., Chen, P., Zheng, H., Gao, X., Xiang, J., & Huang, Y. (2023). Predicting the utilization factor of blasthole in rock roadways by random forest. *Underground Space, 11*, 232–245. https://doi.org/10.1016/j.undsp.2023.01.006

Mohammadi, A., Faraji, M., Ebrahimi, A. A., Nemati, S., Abdolahnejad, A., & Miri, M. (2020). Comparing THMs level in old and new water distribution systems; seasonal variation and probabilistic risk assessment. *Ecotoxicology and Environmental Safety, 192*, 110286. https://doi.org/10.1016/j.ecoenv.2020.110286

Okoji, C. N., Okoji, A. I., Ibrahim, M. S., & Obinna, O. (2022). Comparative analysis of adaptive neuro-fuzzy inference system (ANFIS) and RSRM models to predict DBP (trihalomethanes) levels in the water treatment plant. *Arabian Journal of Chemistry, 15*(6), 103794. https://doi.org/10.1016/j.arabjc.2022.103794

Ozgur, C., Kaplan-Bekaroglu, S. S. (2022). Carbonaceous disinfection by-products in low Suva waters: occurrence, formation potential, and health risk assessment. Applied Ecology and Environmental Research, 20(5), 3833–3851. https://doi.org/10.15666/aeer/2005_38333851

Peng, F., Lu, Y., Wang, Y., Yang, L., Yang, Z., & Li, H. (2023). Predicting the formation of disinfection by-products using multiple linear and machine learning regression. *Journal of Environmental Chemical Engineering, 11*(5), 110612. https://doi.org/10.1016/j.jece.2023.110612

Peng, F., Peng, J., Li, H., Li, Y., Wang, B., & Yang, Z. (2020). Health risks and predictive modeling of disinfection byproducts in swimming pools. *Environment International, 139*, 105726. https://doi.org/10.1016/j.envint.2020.105726

Peng, X., & Chen, D. (2018). PTSVRs: Regression models via projection twin support vector machine. *Information Sciences, 435*, 1–14. https://doi.org/10.1016/j.ins.2018.01.002

Pérez-Lucas, G., Martínez-Menchón, M., Vela, N., & Navarro, S. (2022). Removal assessment of disinfection by-products (DBPs) from drinking water supplies by solar heterogeneous photocatalysis: A case study of trihalomethanes (THMs). *Journal of Environmental Management, 321*, 115936. https://doi.org/10.1016/j.jenvman.2022.115936

Platikanov, S., Martín, J., & Tauler, R. (2012). Linear and nonlinear chemometric modeling of THM formation in Barcelona's water treatment plant. *Science of the Total Environment, 432*, 365–374. https://doi.org/10.1016/j.scitotenv.2012.05.097

Shi, X., Liu, D., Chen, L., Lin, Y., Fu, M.-L., Sun, W., & Yuan, B. (2023). Challenges of point-of-use devices in purifying tap water: The growth of biofilm on filters and the formation of disinfection byproducts. *Chemical Engineering Journal, 462*, 142235. https://doi.org/10.1016/j.cej.2023.142235

Singh, K. P., & Gupta, S. (2012). Artificial intelligence based modeling for predicting the disinfection by-products in water. *Chemometrics and Intelligent Laboratory Systems, 114*, 122–131. https://doi.org/10.1016/j.chemolab.2012.03.014

Srivastav, A. L., Kaur, T. (2020). Chapter 18 - Factors affecting the formation of disinfection by-products in drinking water: human health risk. In M. N. V. Prasad (Ed.), Disinfection By-products in Drinking Water (pp. 433–450). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-08-102977-0.00019-6

Uyak, V., Toroz, I., & Meriç, S. (2005). Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul. *Desalination, 176*(1), 91–101. https://doi.org/10.1016/j.desal.2004.10.023

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-2440-0

Xu, Q., Zhang, J., Jiang, C., Huang, X., & He, Y. (2015). Weighted quantile regression via support vector machine. *Expert Systems with Applications, 42*(13), 5441–5451. https://doi.org/10.1016/j.eswa.2015.03.003

Xu, Z., Shen, J., Qu, Y., Chen, H., Zhou, X., Hong, H., & Wu, F. (2022). Using simple and easy water quality parameters to predict trihalomethane occurrence in tap water. *Chemosphere, 286*, 131586. https://doi.org/10.1016/j.chemosphere.2021.131586

Yang, H., Ye, S., Wang, J., Wang, H., Wang, Z., Chen, Q., & Tan, X. (2021). The approaches and prospects for natural organic matter-derived disinfection byproducts control by carbon-based materials in water disinfection progresses. *Journal of Cleaner Production, 311*, 127799. https://doi.org/10.1016/j.jclepro.2021.127799

Zhang, M., Deng, Y.-L., Liu, C., Lu, W.-Q., & Zeng, Q. (2023). Impacts of disinfection byproduct exposures on male reproductive health: Current evidence, possible mechanisms and future needs. *Chemosphere, 331*, 138808. https://doi.org/10.1016/j.chemosphere.2023.138808

Zheng, W., Tian, D., Wang, X., Tian, W., Zhang, H., Jiang, S., & Qu, W. (2013). Support vector machine: Classifying and predicting mutagenicity of complex mixtures based on pollution profiles. *Toxicology, 313*(2), 151–159. https://doi.org/10.1016/j.tox.2013.01.016

Zheng, Y., Ge, Y., Muhsen, S., Wang, S., Elkamchouchi, D. H., Ali, E., & Ali, H. E. (2023). New ridge regression, artificial neural networks and support vector machine for wind speed prediction. *Advances in Engineering Software, 179*, 103426. https://doi.org/10.1016/j.advengsoft.2023.103426

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.