scientific reports

Check for updates

OPEN

Water quality prediction and carbon reduction mechanisms in wastewater treatment in Northwest cities using Random Forest Regression model

Jingjing Sun[1,6], Xin Guan[2,6], Xiaojun Sun[3,6]✉, Xiaojing Cao[4], Yepei Tan[5] & Jiarong Liao[1]

With the accelerated urbanization and economic development in Northwest China, the efficiency of urban wastewater treatment and the importance of water quality management have become increasingly significant. This work aims to explore urban wastewater treatment and carbon reduction mechanisms in Northwest China to alleviate water resource pressure. By utilizing online monitoring data from pilot systems, it conducts an in-depth analysis of the impacts of different wastewater treatment processes on water quality parameters. This work pays particular attention to their impact on key indicators such as Chemical Oxygen Demand (COD), NH_4^+-N, Total Phosphorus (TP), and Total Nitrogen (TN), and the application of predictive models. The work first establishes a Random Forest Regression (RFR) model. The RFR algorithm integrates Bagging ensemble learning and random subspace theory to construct multiple decision trees and aggregate their predictions, thereby enhancing the model's prediction accuracy and stability. Using bootstrap sampling, the RFR model generates multiple training subsets from the original data and randomly selects subsets of variables to construct regression trees. Its performance in predicting various water quality indicators is then evaluated. The results show that the RFR model exhibits excellent performance, achieving high levels of prediction accuracy and stability for all indicators. For example, the R^2 for COD prediction is 0.99954, while the R^2 values for NH_4^+-N, TP, and TN predictions reach 0.99989. Compared to five other models, the RFR model demonstrates the best performance across all water quality indicator predictions. This work provides critical support for optimizing wastewater treatment technologies and developing water resource management policies. These findings also offer essential theoretical and empirical insights for the future improvement of urban wastewater treatment technologies and water resource management decision-making.

Keywords Random Forest Regression, Wastewater treatment, Water quality indicators, Predictive models, Carbon emission reduction mechanism

With the rapid economic development and accelerated urbanization in China, the Northwest region, as a crucial ecological barrier and resource base for the country, is facing increasingly severe challenges in water resource management. Against the backdrop of mounting water scarcity and pollution issues, alleviating water resource pressures and achieving sustainable management have become pressing problems[1]. Urban wastewater treatment plays a vital role in this process. It serves as a fundamental measure to ensure water environmental safety and public health and acts as a core pathway for reducing carbon emissions and advancing ecological civilization[2]. The unique arid and semi-arid climate conditions and complex terrain of the Northwest region present numerous technical challenges and specific requirements for urban wastewater treatment. Accelerated industrialization and urbanization have further exacerbated water resource pressures, leading to a rapid increase in wastewater

[1]School of Public Administration, Guangzhou University, Guangzhou 510006, China. [2]Guangzhou Xinhua University, Dongguan 523133, China. [3]School of Foreign Languages, Hubei University of Economics, Wuhan 430205, China. [4]Master of Business Administration, London Metropolitan University, London N7 8DB, UK. [5]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China. [6]These three authors considered as co-first authors: Jingjing Sun, Xin Guan and Xiaojing Cao. ✉email: sunxiaojun@hbue.edu.cn

discharge. This necessitates more efficient and energy-saving treatment technologies[3]. Additionally, carbon emissions associated with wastewater treatment have become a critical environmental concern. Achieving synergistic optimization of wastewater treatment and carbon reduction is essential for the region's ecological and socio-economic sustainable development[4].

With growing environmental awareness, an increasing number of wastewater treatment plants have been established. However, they have also become a significant source of greenhouse gas emissions[5]. Bai et al.[6] analyzed the greenhouse gas emission factors of Anoxic/Oxic (A/O) and Sequencing Batch Reactor (SBR) methods in urban wastewater treatment. It was found that the total greenhouse gas emissions of the A/O method (415.63 gCO_2 equivalent/m^3) were significantly lower than those of the SBR method (879.51 gCO_2 equivalent/m^3). The results also highlighted that under specific dissolved oxygen conditions, the N_2O emission factor of the A/O method could be reduced to 0.29% of the influent nitrogen content. Besides, the ammonia oxidation rate in the SBR method was significantly affected by temperature, producing more greenhouse gases at 25 °C. Greenhouse gas emissions are associated with the wastewater treatment process. Dui et al.[7] proposed a multi-stage resilience approach for urban wastewater treatment networks based on phase and node recovery importance, aiming to enhance the resilience of wastewater systems. This approach modeled and evaluated resilience across drainage, treatment, and recovery stages, and used importance metrics to prioritize recovery efforts and enhance system resilience during failure and restoration processes. Marin and Rusănescu[8] studied the application of sludge from wastewater treatment plants in Alexandria, Romania. It was found that sludge rich in organic matter and nutrients could improve soil fertility without exceeding the maximum permissible heavy metal concentrations. Thus, wastewater sludge can serve as fertilizer for degraded soils, alleviating water resource pressures. Su et al.[9] updated the Computable General Equilibrium-System Dynamics Water Environment (CGE-SyDWEM) model to simulate the water-energy-carbon nexus of China's integrated urban drainage systems. They found that wastewater treatment plants in Shenzhen accounted for 89% of total greenhouse gas emissions, and optimizing carbon reduction strategies and water engineering practices could reduce emissions by 7% by 2025. Xian et al.[10] combined Life Cycle Assessment (LCA), Data Envelopment Analysis (DEA), and surveys to evaluate Green House Gas (GHG) emissions associated with wastewater treatment in Shenzhen from 2005 to 2020. They revealed that indirect emissions from sludge treatment were the primary source. Moreover, they highlighted that future wastewater treatment would face a greater potential increase in greenhouse gas emissions, and emphasized the urgent need for innovative environmental management measures and the promotion of water-saving practices. Jiménez-Benítez et al.[11] investigated a semi-industrial-scale Anaerobic Membrane Bioreactor (AnMBR) urban wastewater treatment plant. It was found that the treatment plant could operate at ambient temperatures for 580 days without requiring chemical cleaning and function as a net energy producer for most of the experimental period. This technology, by incorporating degassing membranes, achieved low net energy requirements, demonstrating its potential as an alternative to traditional wastewater treatment methods. Given the unique geographic environment and relative scarcity of water resources in the Northwest region, wastewater treatment and reuse have become critical pathways to alleviate water resource pressures.

In recent years, with the increasing demand for wastewater treatment, Machine Learning (ML)-based methods have gained widespread attention for improving the performance of wastewater treatment plants. The following studies highlight the application and potential of ML technologies in wastewater treatment plants across different regions. Mahanna et al.[12] studied the performance of the AlHayer wastewater treatment plant in Saudi Arabia and used ML techniques to predict key physico-chemical parameters, including Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), and Suspended Solids (SS). They evaluated the performance of four models: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR). The results showed that the RF model performed best in predicting COD and SS, with Coefficient of Determination (R^2) values of 91% and 95%, respectively. The GB model performed best for predicting BOD, with an R^2 of 92%. This indicated that RF and GB models had significant advantages in estimating the physico-chemical characteristics of wastewater treatment plants and could support the optimization of treatment processes. Zhang et al.[13] explored the impact of microbial community structure in activated sludge systems on wastewater treatment performance and used ML models to predict phosphorus and nitrogen removal efficiencies. They performed a meta-analysis of high-throughput sequencing data and identified key microbial genera associated with phosphorus and nitrogen removal. Among the six ML models, Extreme Gradient Boosting (XGBoost) demonstrated the highest prediction accuracy. Additionally, the study identified 13 key microbial genera through cross-entropy, which played important roles in the phosphorus and nitrogen cycles. The results emphasized the potential of combining microbial data and ML technologies in wastewater treatment plant design and provided a new approach for optimizing biological processes in wastewater treatment. Cechinel et al.[14] focused on developing ML models to predict the COD concentration of effluents from wastewater treatment plants. They tested Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), and RF models using a dataset from the Umbilo wastewater treatment plant in South Africa. The study found that MLP performed best for COD prediction on daily datasets, while LSTM performed better on hourly datasets, making it suitable for handling high-frequency data with time series information. Variable importance analysis showed that Total Suspended Solid (TSS) was a key variable for predicting COD. Their study demonstrated the potential application of ML models in optimizing wastewater treatment processes and emphasized the importance of validating models with real measurement data. Rios Fuck et al.[15] examined the applicability of ML models in predicting wastewater quality parameters and explored the impact of plant operation changes on model performance. They evaluated the performance of RF, SVM, and MLP models in both simulated and real-world scenarios. The results indicated that the RF model adapted well to real data from the Ambev wastewater treatment plant, while the MLP model had higher accuracy in predicting Total Nitrogen (TN) in the simulated Water Evaluation and Simulation Tool (WEST) scenario. Through Partial Dependence Plot (PDP) and Permutation Importance (PI) analysis, the study revealed key inflow parameters

related to nitrogen content. Their research highlighted the importance of high-quality data and dynamic operational information and provided new insights for optimizing wastewater treatment plant performance.

Overall, although significant progress has been made in wastewater treatment performance prediction, further exploration is needed in areas such as regional applicability of data, quantification of microbial communities, multi-objective optimization, and model interpretability. This work aims to address these gaps and promote the innovative application of ML in the wastewater treatment field. It intends to provide new theoretical and practical support for improving wastewater treatment efficiency and achieving sustainable development goals. Specifically, this work reviews the current status of urban wastewater treatment systems in the northwest region, and analyzes the impact of water resource shortages and pollution on regional ecology and economy. Moreover, it investigates the purification effects of key water quality parameters such as COD, NH_4^+-N, Total Phosphorus (TP), and TN and the carbon emission influence. Then, this work optimizes the wastewater treatment system based on RF algorithms and field data. It also proposes technical and management strategies to support regional sustainable development. This work is expected to provide theoretical foundations and practical guidance for the technological innovation and management optimization of wastewater treatment systems in the northwest region, contributing to sustainable water resource use and ecological environmental improvement.

Data-driven urban wastewater treatment

Current status of wastewater discharge

Figure 1 illustrates the national wastewater discharge data for 2023.

In 2023, among the major cities in Northwest China, Xinjiang exhibited the highest COD in industrial wastewater, reaching 10,500 tons. It indicates a significant burden in industrial wastewater treatment in the region. Inner Mongolia ranked second with a COD of 6700 tons. Gansu and Qinghai followed in third and fourth place, with COD values of 3500 tons and 2000 tons, respectively. The data suggest significant differences in industrial wastewater treatment across different cities, likely influenced by variations in industrial development levels and pollution control measures.

Figure 2 presents the comprehensive water environment index for Northwest China.

In recent years, the water environment index in Northwest China has shown a gradual improvement trend. This progress is attributed to the government's efforts to strengthen water pollution control, promote ecological civilization construction, and implement a series of environmental protection policies and measures. However, due to the relative scarcity of water resources and the fragile ecological environment in the region, water environment protection and management continue to face numerous challenges.

Figure 3 illustrates the water functional zones in Northwest China.

According to data released by the Ministry of Ecology and Environment, the proportion of surface water monitoring sections with good water quality (Class I–III) in Northwest China has been steadily increasing in recent years. For instance, the 2023 report indicates that the proportion of good water quality in the region's rivers remains at a high level. This demonstrates a continuous improvement in water quality conditions. Meanwhile,

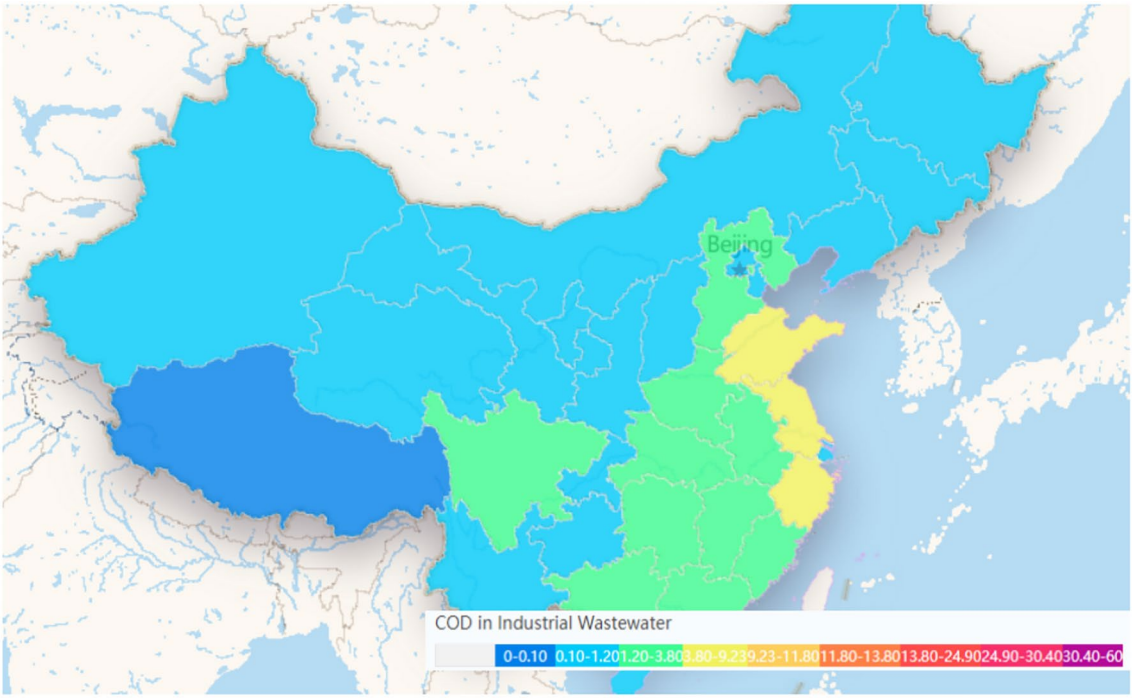

Fig. 1. National wastewater discharge data (Note: The data was obtained from the Institute of Public & Environmental Affairs (IPE): http://www.ipe.org.cn/. The information about license can be found below: https://wwwen.ipe.org.cn/about/disclaimer.html.).

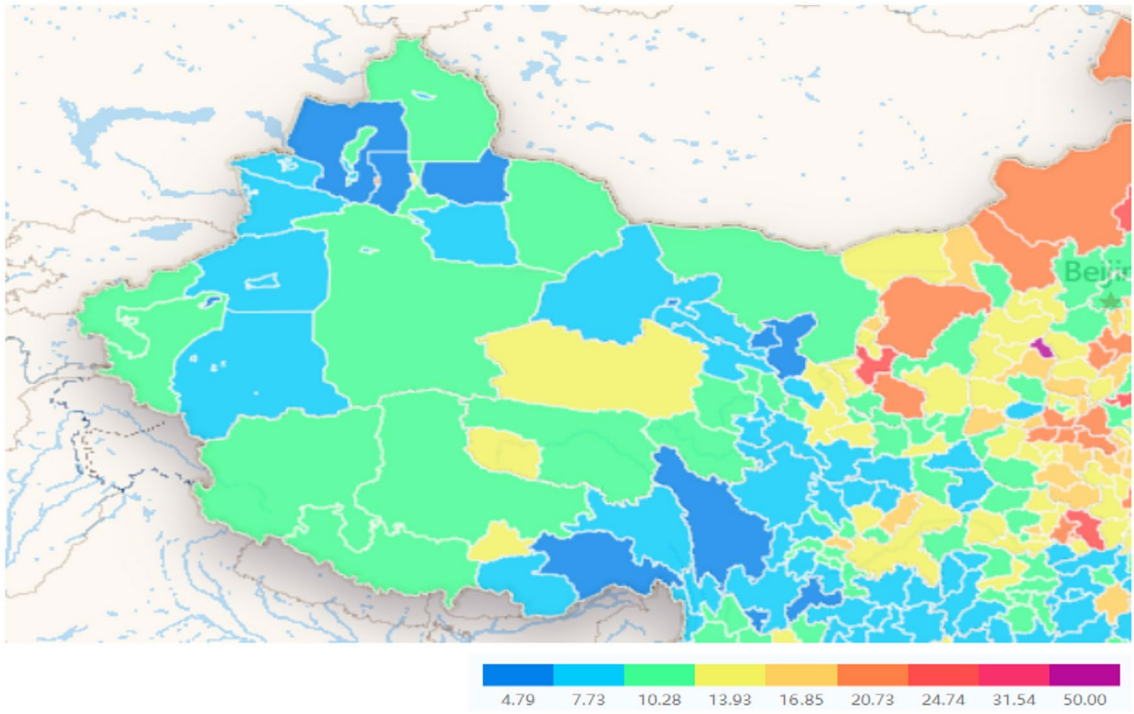

Fig. 2. Comprehensive water environment index in northwest China (Note: The data was obtained from the Institute of Public & Environmental Affairs (IPE): http://www.ipe.org.cn/. The information about license can be found below: https://wwwen.ipe.org.cn/about/disclaimer.html.).

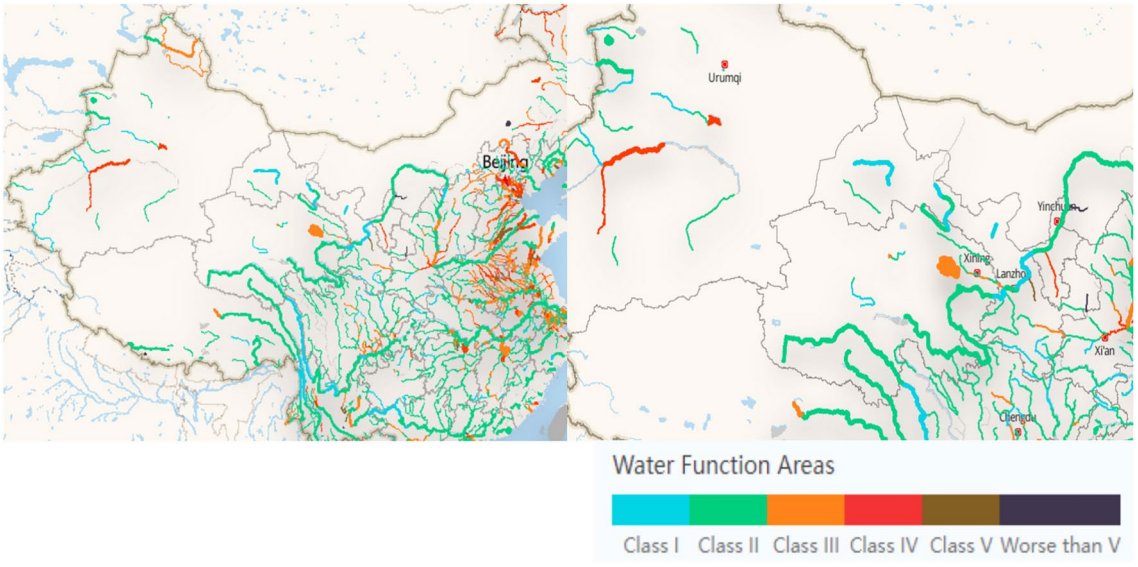

Fig. 3. Water functional zones in northwest China (Note: The data was obtained from the Institute of Public & Environmental Affairs (IPE): http://www.ipe.org.cn/. The information about license can be found below: https://wwwen.ipe.org.cn/about/disclaimer.html.).

the proportion of Class V sections (representing the worst water quality) has also been declining year by year. This signifies that Northwest China has made significant progress in eliminating severe water pollution.

Random Forest Regression algorithm

This work employs the Random Forest Regression (RFR) model to analyze and predict the impact of water quality indicators. It is based on the following principles. (1) Ensemble Learning: The RFR model is a regression method based on ensemble learning. It constructs multiple decision trees and integrates their predictions to

mitigate the risk of overfitting that may arise from a single decision tree and enhance the model's generalization ability. By combining the predictions of multiple trees, RFR can more accurately capture complex patterns and nonlinear relationships within the data. (2) High Fault Tolerance: RFR uses a strategy of randomly selecting data subsets and feature subsets, training each decision tree on a unique combination of samples and features. This increases model diversity and ensures that even if certain data points introduce errors, the overall prediction remains highly stable. (3) Noise Resistance: RFR is robust to noisy data. When dealing with datasets containing significant noise, the integration of multiple trees helps reduce the impact of noise on the final prediction results. (4) Feature Importance Evaluation: RFR can automatically compute the importance of each feature in the prediction process and provide an intuitive ranking of feature importance. This capability aids in identifying the most critical features for predicting the target variable, thereby optimizing the model further. (5) Nonlinear Relationship Handling: Unlike traditional linear regression models, RFR is capable of addressing complex nonlinear relationships. This adaptability makes it particularly suitable for multidimensional, nonlinear features encountered in real-world problems, such as modeling complex systems like wastewater treatment.

The RFR algorithm is an ensemble method developed from the decision tree model. It integrates Bagging ensemble learning with the random subspace method. By randomly selecting subsets of variables for each split, it ensures diversity among the decision trees. Ultimately, by combining the predictions of all the trees, the overall prediction variance is reduced, and the model's accuracy is improved[16]. Depending on the type of dependent variable, the random forest algorithm can be categorized into classification and regression models. The random forest classification model is used for classification problems, while the RFR model is employed for regression problems. The RFR model utilizes the bootstrap resampling method to randomly select a subset of samples from the dependent variable's dataset and a subset of variables from the independent variables. They serve as the nodes of the regression trees. This approach ensures differences among the constructed regression trees[17]. Typically, a random forest consists of hundreds or even more regression trees, and the final result is determined by aggregating the predictions of all the regression trees.

The RFR algorithm predicts by combining multiple regression tree models. The equation for calculating the model's predicted value is as follows:

$$\overline{y} = \frac{1}{j} \cdot \sum_{j=1}^{m} h\left(X, \theta_j\right) \tag{1}$$

$\overline{y}$ represents the average prediction value of all the regression trees; j denotes the number of regression trees, where $j = (1, m)$; $h\left(X, \theta_j\right)$ is the prediction value of the j-th regression tree for the input X.

As the number of regression trees j approaches infinity, the regression function of the model can be expressed as:

$$E_{X,Y}\left(Y - average_j h\left(X, \theta_j\right)\right)^2 \rightarrow E_{X,Y}\left(Y - E_\theta h\left(X, \theta\right)\right)^2 \tag{2}$$

In the operation of the RFR model, the expression $Y = average_j h\left(X, \theta_j\right)$ as $j \rightarrow +\infty$ is typically used to approximate the model's regression function $Y = E_\theta h\left(X, \theta\right)$. In this context, the generalization error of any single regression tree predictor is $E_{X,Y}\left(Y - h\left(X\right)\right)^2$. The average generalization error of the RFR algorithm is represented as:

$$PE^*\left(tree\right) = E_\theta E_{X,Y}\left(Y - h\left(X, \theta\right)\right)^2 \tag{3}$$

It is assumed that the relationship $E\left(Y\right) = E_x h\left(X, \theta\right)$ holds for all θ. Then, it can be obtained that:

$$PE^*\left(forest\right) \leq \overline{\rho} \cdot PE^*\left(tree\right) \tag{4}$$

$\overline{\rho}$ represents the correlation coefficient between $Y - h\left(X, \theta\right)$ and $Y - h\left(X, \theta'\right)$, where θ and θ' are independent of each other. The generalization error of the random forest is $\overline{\rho}$ times the generalization error of a single regression tree. By introducing θ and θ', the RFR model's accuracy is enhanced.

Establishment process of the RFR algorithm

The selected dataset comes from the online monitoring data of a pilot system. These data are collected through a combination of in-situ detection and ex-situ monitoring operations. The in-situ detection system is applied to obtain relevant data that are easy to monitor within the pilot system, such as real-time data on dissolved oxygen, pH, oxidation–reduction potential, and temperature. The ex-situ monitoring system is adopted to monitor the water quality parameters of the pilot system. This system uses a water sample collection pump to draw sewage from the secondary sedimentation tank of the pilot system into a storage tank for water quality monitoring. The main monitoring indicators include COD, NH_4^+-N, TN, and TP[18,19]. The purpose of these data is to predict the impact of influent indicator concentrations and system operating parameters on effluent indicator concentrations. Figure 4 illustrates the process of generating the forest in the RFR model.

The RFR model framework consists of several key modules. First is the input module, also known as the training sample input module. In this stage, the model receives and processes training data collected from the pilot system's online monitoring. The data include various variables and indicators affecting the wastewater treatment process, such as water quality parameters (COD, NH_4^+-N, TN, and TP) and process parameters (dissolved oxygen, pH, oxidation–reduction potential, and temperature). Next is the random sampling module, where the random forest algorithm uses Bootstrap resampling to randomly draw multiple subsets from the training

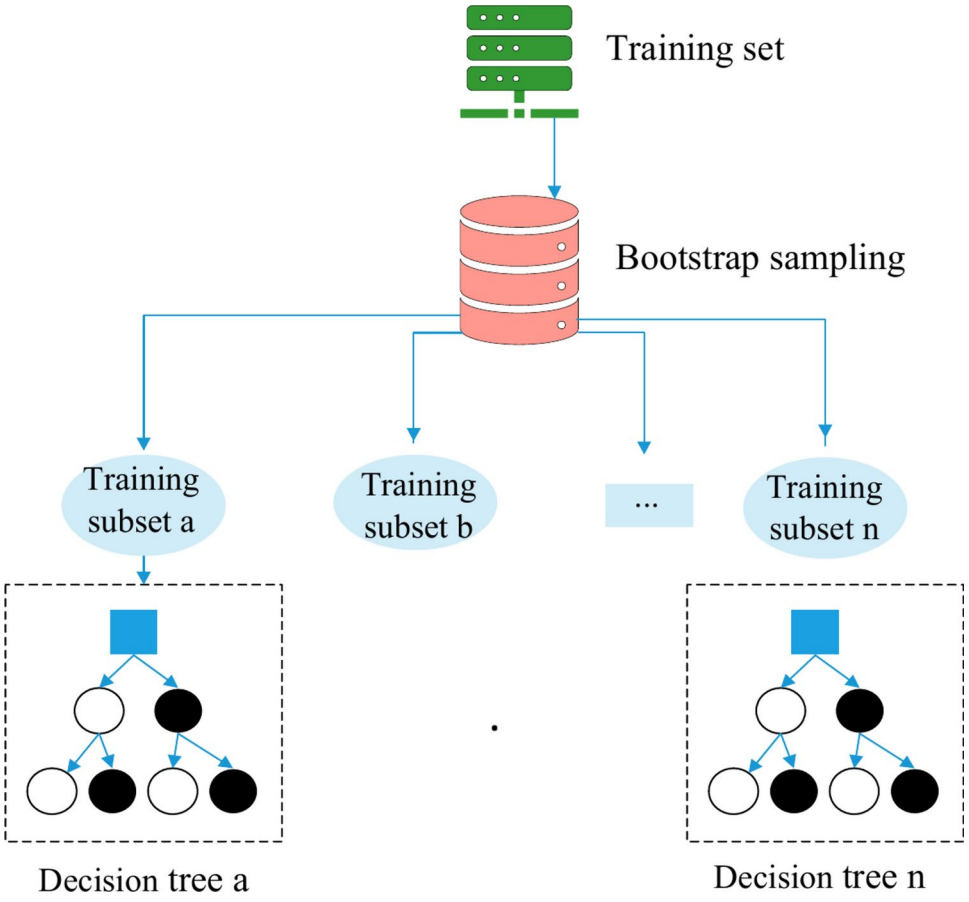

Fig. 4. Process of generating the forest in the RFR model.

samples. Each subset is used to construct an independent regression tree, ensuring diversity and preventing overfitting, thus improving the model's generalization capability. Then, there is the random forest module, which is a regression forest constructed from numerous regression trees. In this module, each regression tree is built from the subsets generated by the random sampling module[20]. During the construction of each regression tree, the random forest algorithm randomly selects a portion of the optimal variables for splitting, further increasing the differences between the trees. Ultimately, the predictions of all the regression trees are aggregated using ensemble learning by averaging the predictions of all the trees to obtain the final model prediction. Figure 5 illustrates the prediction process of the RFR model.

The output module, representing the model's prediction results, is the final component of the RFR model. In this module, the data samples prepared through preprocessing are input into the RFR model. The preprocessing involves the following steps. (1) Data Cleaning: Missing values, outliers, and duplicate data are removed. Missing values are addressed using methods such as interpolation or mean imputation to ensure dataset completeness. Outliers are identified using a standard deviation-based rule, where data points outside a reasonable range are treated appropriately. (2) Data Normalization: Due to the differing scales of various features, data standardization is performed to prevent certain features from excessively influencing model training. This involves methods such as Z-score standardization, where each feature is adjusted by subtracting its mean and dividing by its standard deviation. This transforms the data to have a mean of 0 and a standard deviation of 1. (3) Data Splitting: The dataset is divided into training and testing sets, with 70%-80% of the data used for training and the remaining 20%-30% reserved for testing. This ensures the model's performance is evaluated on unseen data, reducing the risk of overfitting.

Next, preprocessed data samples are fed into the RFR model. Here, the data samples first pass through the Bootstrap Sampling module. In this module, Bootstrap resampling techniques are used to randomly extract multiple subsets from the original training data. These subsets are used to construct multiple independent regression trees, each representing different parts and characteristics of the overall data. Subsequently, the data samples enter the random forest module[21]. In this module, the random forest algorithm randomly selects optimal variables for splitting, constructing numerous regression trees. Introducing randomness during tree construction ensures each tree has distinct characteristics, thereby enhancing the model's stability and generalization capability. Each regression tree learns from and trains on the input data samples, generating its own predictions[22]. Finally, the predictions from all regression trees are aggregated and processed in the output module. Specifically, the final prediction of the RFR model is derived by calculating the average of all regression tree predictions. This average represents the model's prediction of effluent indicators in wastewater treatment

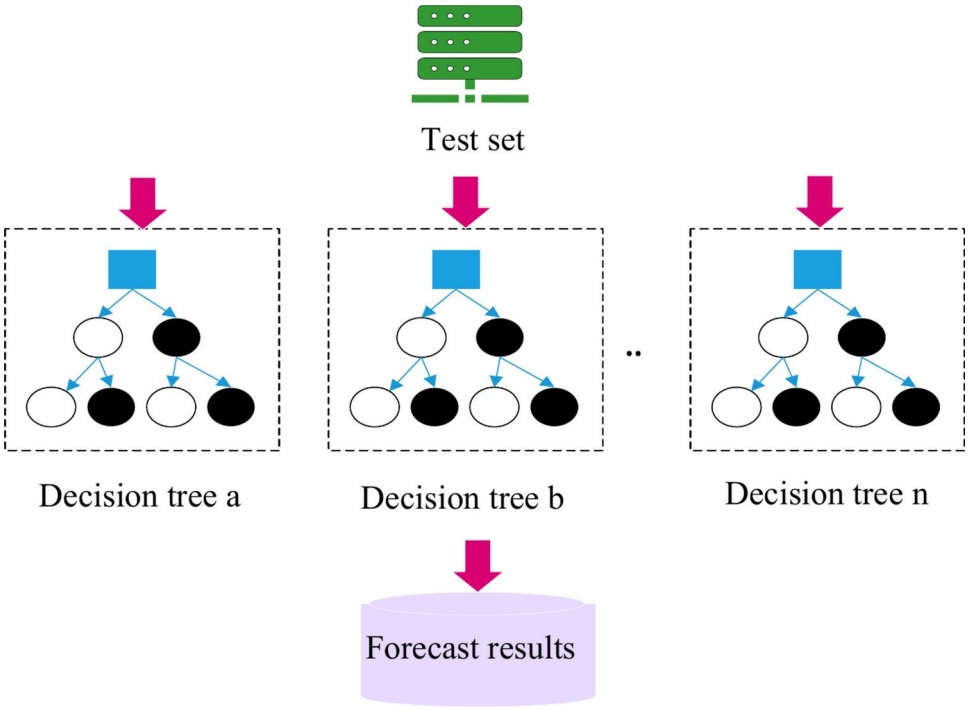

Fig. 5. Prediction process of the RFR model.

systems. Through this ensemble learning method, the RFR model effectively reduces prediction variance, thereby improving overall prediction accuracy and reliability.

Through this comprehensive process, the RFR model accurately predicts effluent indicators in wastewater treatment systems and enhances its adaptability and predictive performance for complex systems through continuous learning and optimization. These predictions provide crucial insights for the operational management of wastewater treatment plants. This aids in process optimization, reducing environmental impact, and enhancing the efficiency and effectiveness of wastewater treatment. Overall, the RFR model achieves effective modeling and prediction of wastewater treatment processes through the collaborative efforts of these modules. The input module ensures the diversity and comprehensiveness of training data, while the random sampling and random forest modules enhance model stability and accuracy by introducing randomness and diversity[23]. This framework effectively addresses the complexity and non-linearity of wastewater treatment processes and serves as a scientific basis for optimizing systems in practical operations.

Training process of the RFR algorithm

Bootstrap sampling is a commonly used technique in non-parametric statistics that involves sampling with replacement. Its essence lies in repeatedly sampling from the training data to learn the overall distribution trends and perform statistical inference. Specifically, in the context of wastewater treatment, influent concentrations such as COD, NH_4^+-N, and TP are represented over time as $X_{COD}(t)$, $X_{NH_4}(t)$ and $X_{TP}(t)$, $t=1, 2, 3, \ldots, n$. Other variables such as aeration in the aerobic tank, internal and external recirculation rates, and sludge discharge are denoted as $X_{O_2}(t)$, $X_{TR}(t)$, and $X_S(t)$, respectively. These variables constitute the input variables of the model, that is, the independent variables. The independent variable X is encoded and input into the Bootstrap sampling module to form training subsets, denoted as X_{train}. Subsequently, the training subset X_{train} is fed into the random forest module and further processed to form X_{forest} after encoding. Within the random forest module, model output is obtained through computation according to Eq. (4). Figure 6 illustrates the training process of the RFR algorithm.

The random forest module serves as the core of the algorithm, operating through the following steps: (1) Bootstrap Sampling: it involves randomly selecting n training samples from the initial dataset using the Bootstrap method. This process creates multiple training subsets, with the remaining samples serving as the testing set for each iteration. (2) Regression Tree Construction: it entails selecting m variables (where $m<p$ and p represents the total number of variables, typically set to $m = \sqrt{p}$) at each branch node to split candidates randomly. The optimal split is determined based on Mean Squared Error (MSE). (3) Regression Tree Growth: it involves the independent growth of each regression tree from the root node to the internal node, to the leaf nodes[24]. The growth process stops based on preset conditions, specifically the maximum depth d of the tree. The predictions from all regression trees are aggregated and their average is computed to derive the final prediction of the RFR model. Model accuracy is evaluated using Mean Absolute Error (MAE), MSE, and R^2, metrics that objectively reflect the model's precision and provide effective guidance for optimizing and operating subsequent wastewater treatment processes.

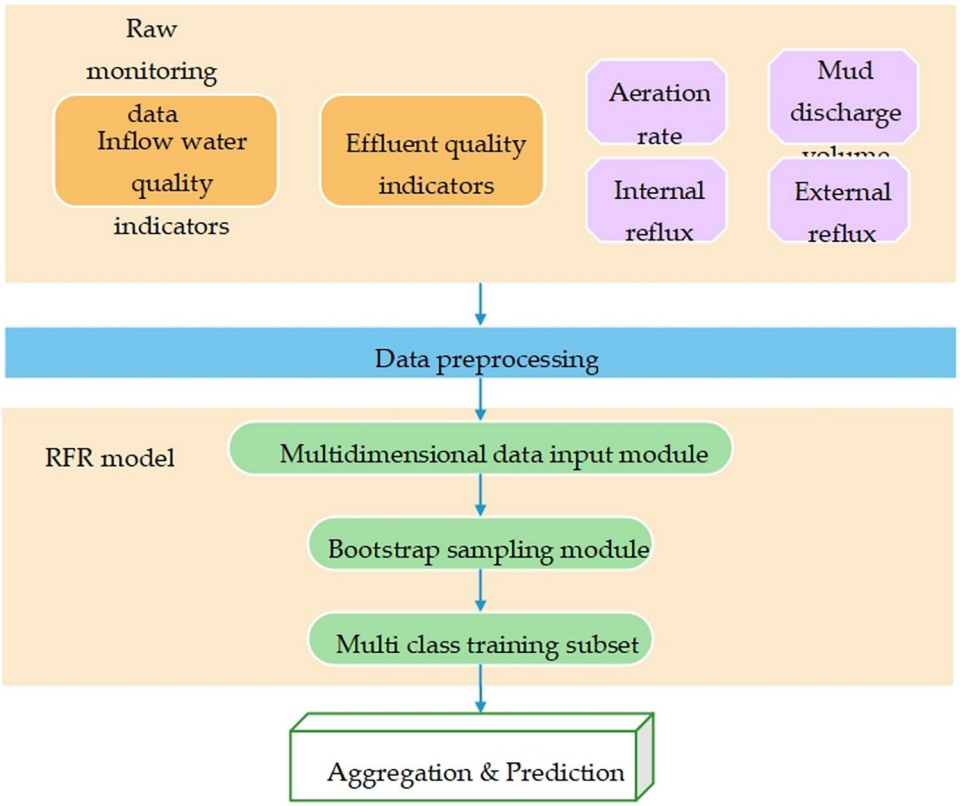

Fig. 6. RFR algorithm training process.

Model evaluation metrics

Prediction accuracy plays a crucial role in carbon reduction strategies for wastewater management. The wastewater treatment process involves significant energy consumption, particularly during biological and chemical treatment stages. Accurately predicting key parameters, such as pollutant concentration, treatment efficiency, and effluent quality, helps optimize treatment processes, minimize energy waste, and reduce carbon emissions. Reliable predictive models enable operators to anticipate potential issues in the treatment process, such as excessive chemical dosing or unnecessary energy demand, and take corrective actions accordingly. By accurately forecasting energy usage at different treatment stages, managers can optimize energy utilization while meeting environmental standards, thereby minimizing unnecessary emissions. Carbon reduction targets often need to be set based on specific treatment capacities and operational efficiencies. Without accurate prediction capabilities, excessive treatment demands may lead to resource wastage and increased carbon emissions. Accurate predictions ensure that the treatment process aligns with actual requirements, avoiding overdesign or excessive input, and ultimately contributing to efficient and sustainable wastewater management.

In order to objectively assess the performance of the RFR model, evaluation metrics are used to measure the model's predictive accuracy on the test set. This work primarily employs MAE, MSE, and R^2 as evaluation standards. Below are the equations for calculating each evaluation metric:

MAE: it is to calculate the average absolute difference between the model's predicted values and the actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{5}$$

y_i represents the actual value, $\widehat{y_i}$ represents the predicted value, and n is the number of samples.

MSE: it is to calculate the average of the squared differences between predicted values and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{6}$$

R^2: it is to measure the proportion of the total variance explained by the model's explanatory variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{7}$$

These evaluation metrics provide different perspectives for assessing model performance. The MAE directly reflects the average absolute difference between predicted and actual values, while the MSE emphasizes the impact of larger errors. R^2 demonstrates the model's ability to explain the variability in the data. MAE and Root Mean Squared Error (RMSE) values approaching 1 indicate larger errors in the model predictions, suggesting lower performance. Conversely, smaller MAE and RMSE values indicate better model performance. In contrast, a higher R^2 value indicates better model performance as it reflects greater explanatory power of the model over the data variability.

In water treatment prediction, MAE provides a straightforward measure of error, reflecting the magnitude of the model's prediction bias. For instance, when predicting a water quality parameter such as COD, a small MAE indicates that the model's predictions are very close to the actual measurements, signifying high predictive accuracy. MAE helps evaluate whether the model can provide sufficiently precise guidance in real-world operations to optimize treatment processes and resource allocation. The operation of water treatment facilities relies on accurate water quality predictions to ensure effective treatment, particularly for real-time monitoring and process adjustments. A lower MAE ensures that operators can promptly fine-tune the treatment process, avoiding fluctuations in water quality standards and enhancing system stability. MSE highlights the impact of large errors on the model's overall performance. If certain outliers or extreme cases cause an increase in MSE, it indicates significant errors in the model's handling of such scenarios, which may negatively affect optimization strategies for the treatment system. For example, during peak loads or abnormal pollution events, a high MSE suggests that the model fails to accurately predict these unusual variations. Water treatment facilities often face unexpected events or irregular pollution sources, and the MSE value reflects the model's ability to address these complex situations. A lower MSE ensures stable predictive performance across various operating conditions, enabling decision-makers to take necessary emergency measures swiftly in response to water quality anomalies. A higher R^2 value indicates that the model fits the water quality data well, demonstrating a strong correlation between the predicted and actual values. For example, an R^2 value close to 1 means the model accurately predicts various water quality parameters, such as COD, BOD, and ammonia nitrogen, leading to more precise control of the treatment processes. Fine-grained management of water treatment requires accurate predictions to adjust operations and ensure that effluent water quality meets regulatory standards. A high R^2 value signifies that the model effectively captures the patterns in water quality variations, aiding operators in making more accurate scheduling and control decisions.

Thus, MAE, MSE, and R^2 are not only metrics for evaluating model performance but also critical tools for practical operations. They enable water treatment facilities to assess the reliability and stability of models under different operating conditions, ensuring efficient water quality control and resource management. Lower MAE and MSE values, coupled with a higher R^2 value, indicate that the model delivers high-precision predictions. They can provide strong support for process optimization, reduce energy waste, and ensure that effluent water quality complies with environmental standards.

Results analysis based on northwestern engineering data

Evaluation results of the RFR model

This work intends to comprehensively study and analyze the urban wastewater treatment processes in Northwest China and their impact on alleviating water resource pressure and carbon reduction mechanisms. It utilizes online monitoring data from pilot systems. These data are collected through a combination of in-situ and off-site monitoring, covering key water quality parameters during the wastewater treatment process. This work calculates the MAE, MSE, and R^2 of the RFR model's predictive results for different effluent indicators. Figure 7 illustrates the evaluation results of the RFR model.

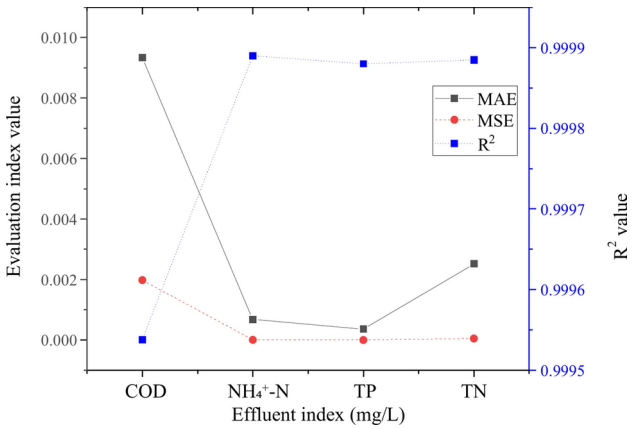

Fig. 7. Evaluation results of the RFR model.

Figure 7 illustrates that the RFR model demonstrates exceptional accuracy and stability in predicting effluent indicators in wastewater treatment. For COD prediction, the RFR model achieves outstanding performance with an MAE of 0.00934, an MSE of 0.00198, and an R^2 of 0.99954. In predicting NH_4^+-N effluent levels, the model excels with an MAE of just 6.76E−4, an MSE of 1.896E−6, and an R^2 of 0.99989. For TP indicators, the RFR model also performs remarkably well, achieving an MAE of 0.000357, an MSE of 5.4E−7, and an R^2 of 0.99988. Similarly, in TN effluent prediction, the model demonstrates excellent performance with an MAE of 0.00251, an MSE of 4.1496E−5, and an R^2 of 0.99989. These results indicate that the RFR model effectively handles and predicts multiple key water quality parameters, providing reliable data support for optimizing and managing wastewater treatment processes.

Model results comparison and analysis

To evaluate the predictive capabilities of different regression models across various indicators, this section compares the performance of the RFR model with five other commonly used regression models: SVR, Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Decision Tree Regression (DTR), and k-Nearest Neighbors Regression (KNNR). Figure 8 presents the performance comparison of these models for each indicator.

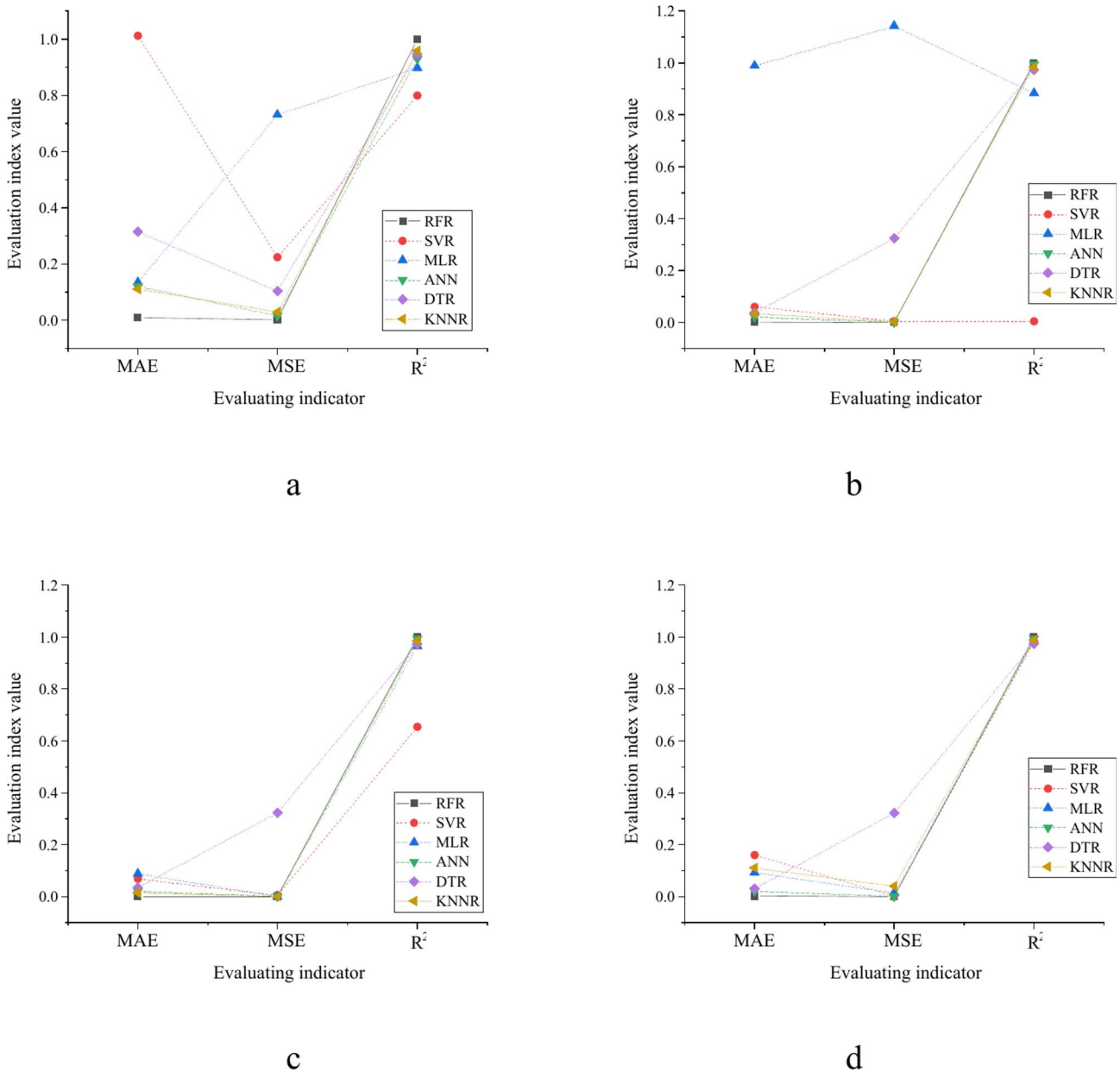

Fig. 8. Performance comparison of each model across different indicators (a: COD indicator; b: NH_4^+-N indicator; c: TP indicator; d: TN indicator).

Figure 8a suggests that the RFR model performs exceptionally well in COD prediction, with an MAE of 0.00934, MSE of 0.00198, and R^2 of 0.99954, demonstrating extremely high prediction accuracy and stability. The SVR model performs well in COD prediction, with an MAE of 1.01123, MSE of 0.22385, and R^2 of 0.79854, showing relatively high prediction accuracy and stability. The MLR model shows moderate predictive capability, with an MAE of 0.13467, MSE of 0.73077, and R^2 of 0.89732. This indicates its high prediction accuracy but lower than that of the RFR and SVR models. The ANN model performs well in COD prediction, with an MAE of 0.1209, MSE of 0.01544, and R^2 of 0.92967, demonstrating high prediction accuracy and stability. The DTR model performs rather averagely, with an MAE of 0.31457, MSE of 0.10379, and R^2 of 0.94601. It shows lower prediction accuracy and higher instability. The KNNR model performs well in COD prediction, with an MAE of 0.11079, MSE of 0.02988, and R^2 of 0.95876, showing high prediction accuracy and stability.

Figure 8b suggests that the RFR model performs exceptionally well in predicting the outflow NH_4^+-N indicator, with an MAE of just 6.76E−4, MSE of 1.896E−6, and R^2 of 0.99989. This indicates that the model's predictions are very close to the actual values, with extremely high prediction accuracy and stability. In comparison, other models such as SVR and MLR have higher MAE and MSE, at 0.06033 and 0.00437, and 1.14189 and 0.88281, respectively. They show lower prediction accuracy and stability than the RFR model. Although the ANN, DTR, and KNNR models perform relatively well, they still cannot match the RFR model. In summary, the RFR model demonstrates excellent accuracy and stability in predicting the outflow NH_4^+-N indicator and is well-suited for water quality prediction and management applications in complex environments.

Figure 8c reveals that the RFR model performs excellently across all evaluation metrics. The RFR model shows outstanding performance, with an MAE of 0.000357, MSE of 5.4E−7, and R^2 of 0.99989, demonstrating extremely high prediction accuracy and stability. The SVR model performs relatively poorly for the TP indicator, with an MAE of 0.06843, MSE of 0.00534, and R^2 of 0.65321, indicating low prediction accuracy and high instability. The MLR model demonstrates moderate predictive ability, with an MAE of 0.08882, MSE of 1.88698E−4, and R^2 of 0.96443, showing high prediction accuracy but still lower than the RFR model. The ANN model performs excellently in TP prediction, with an MAE of 0.02072, MSE of 7.1623E−4, and R^2 of 0.99453, showing extremely high prediction accuracy and stability. The DTR model performs fairly, with an MAE of 0.03452, MSE of 0.32252, and R^2 of 0.97252, indicating lower prediction accuracy and higher instability. The KNNR model performs well in TP prediction, with an MAE of 0.01437, MSE of 5.3277E−4, and R^2 of 0.98429, demonstrating relatively high prediction accuracy and stability.

Figure 8d demonstrates that the RFR model performs outstandingly in predicting the outflow TN indicator, with an MAE of 0.00251, MSE of 4.1496E−5, and R^2 of 0.99989, showing extremely high prediction accuracy and stability. The SVR model performs well for the TN indicator, with an MAE of 0.16022, MSE of 0.00241, and R^2 of 0.98252, indicating relatively high prediction accuracy and stability. The MLR model shows moderate predictive ability, with an MAE of 0.09288, MSE of 0.01424, and R^2 of 0.98421, showing high prediction accuracy but still lower than the RFR and SVR models. The ANN model performs excellently in TN prediction, with an MAE of 0.02072, MSE of 7.1623E-4, and R^2 of 0.99453, demonstrating extremely high prediction accuracy and stability. The DTR model performs fairly, with an MAE of 0.03097, MSE of 0.32242, and R^2 of 0.97422, indicating lower prediction accuracy and higher instability. The KNNR model performs the worst in TN prediction, with an MAE of 0.11097, MSE of 0.03988, and R^2 of 0.98843, indicating relatively high prediction error and lower prediction accuracy.

The RFR model demonstrated the best performance in predicting various water quality indicators, standing out significantly compared to the other five common regression models. By leveraging the idea of ensemble learning, RFR builds multiple decision trees and averages their predictions, effectively reducing the risk of overfitting and handling nonlinear relationships between features well. It possesses strong noise resilience and stability, making it highly accurate in complex wastewater treatment datasets. Although RFR is efficient in handling large datasets, its training process can be time-consuming, and interpreting the specific impact of model features presents some challenges. In contrast, while the SVR can address nonlinear problems and is suitable for small datasets, it requires longer training times and is sensitive to noise and outliers, which affects its stability in complex datasets. The MLR model performs well with linear relationships, but due to its assumption of linearity between features, it struggles with highly nonlinear data and is sensitive to outliers. ANN excels at handling complex nonlinear issues, especially in high-dimensional data, but its training process is computationally intensive, and understanding its internal mechanisms is challenging. Although the DTR model is easy to understand and can handle nonlinear problems, a single tree is prone to overfitting and is sensitive to noisy data. The KNNR model is simple and effective for nonlinear relationships but is limited when dealing with large datasets or high dimensions and is sensitive to feature scaling. Overall, the RFR model stands out due to its powerful nonlinear modeling capability and superior stability, making it the most prominent performer.

Analysis of carbon emissions from wastewater treatment plants

In the wastewater treatment process, the reduction of carbon emissions is a key indicator for evaluating the effectiveness of new methods. To assess the impact of the proposed method on carbon emissions during wastewater treatment, a mass balance approach is used to calculate the carbon equivalent (unit: tons CO_2e) in the treatment process. This work conducts a detailed analysis of factors such as energy consumption, chemical usage, equipment efficiency, and treatment capacity in various stages of the wastewater treatment process, resulting in the calculation of carbon emissions for each stage. Then, by incorporating the optimization measures of the new method, the reduction in carbon emissions for each stage and the overall process is calculated. Table 1 presents a comparison of the carbon emission data for the wastewater treatment plant under the existing and new methods.

Table 1 shows that after adopting the new method, the overall carbon emissions from the wastewater treatment plant decrease by 780 tons of CO_2e per year, resulting in a reduction of 17.33%. The reduction in emissions is also significant at each treatment stage, particularly in the wastewater pretreatment and biological treatment

Treatment stage	Carbon emission (Existing Method) (tons CO_2e/year)	Carbon emission (New Method) (tons CO_2e/year)	Reduction (tons CO_2e/year)	Emission reduction (%)
Wastewater pretreatment	1200	950	250	20.83%
Biological treatment	1500	1200	300	20.00%
Sedimentation & filtration	800	720	80	10.00%
Sludge treatment	1000	850	150	15.00%
Total emissions	4500	3720	780	17.33%

Table 1. Comparison of carbon emission data between the existing and new methods.

stages, where the emission reductions are 20.83% and 20.00%, respectively. This indicates that the new method offers a clear advantage in reducing carbon emissions. These data demonstrate that the new approach not only improves the efficiency of wastewater treatment but also significantly reduces carbon emissions, contributing to the industry's shift toward greener and lower-carbon practices. These results provide a practical, quantified basis for carbon reduction strategies in wastewater management and offer valuable insights for policymakers, environmental agencies, and wastewater treatment plant operators.

Discussion

This work systematically evaluates the application of different regression models in wastewater treatment. It particularly focuses on comparing the performance of the RFR model against five other common regression models (SVR, MLR, ANN, DTR, KNNR). The research findings indicate that the RFR model excels in both prediction accuracy and stability. Particularly in the prediction of water quality indicators such as COD, NH_4^+-N, TP, and TN, it consistently demonstrates lower prediction errors and higher R^2 values. This suggests that the RFR model is highly effective in handling complex nonlinear data relationships, with strong noise resistance and robustness. This makes it suitable for various practical applications in the wastewater treatment process.

Compared to existing research in the literature, the findings of this work further validate the potential of ML models, especially ensemble learning methods, in wastewater treatment. For example, Mahanna et al.[12] 's study demonstrated significant predictive accuracy of RFR in the AlHayer wastewater treatment plant in Saudi Arabia. Moreover, Zhang et al.[13] explored the impact of microbial communities on wastewater treatment through the XGBoost model. Both studies highlighted the potential of machine learning technologies in enhancing the performance of wastewater treatment plants. By comparing the RFR model with other common regression models, this work further affirms the advantages of ensemble learning methods in practical operations.

In real-world applications, the findings of this work hold significant implications for policymakers, operators, and environmental managers in the wastewater treatment industry. First, policymakers can utilize the efficient predictive models proposed to optimize wastewater treatment policies, particularly in predicting and managing key water quality indicators in sewage. Accurately forecasting these indicators helps to implement more precise control measures in real-world operations, reducing environmental pollution and improving treatment efficiency. Moreover, operators of wastewater treatment plants can adjust their operational strategies and optimize treatment processes based on the predictions made by these models, reducing energy consumption and operational costs. For example, the RFR model demonstrates strong capabilities in predicting multiple water quality parameters, enabling operators to identify potential issues earlier and take preventive actions. This not only improves treatment efficiency but also ensures the stable operation of the system under varying environmental conditions. Additionally, stakeholders, including environmental protection agencies and the public, can benefit from the results of this work. By accurately predicting the pollutant concentrations in sewage, relevant authorities can implement more scientific monitoring and management measures. This can ensure that water quality meets environmental standards while minimizing the negative impact of wastewater treatment plants on the environment. With the growing global focus on water resources and environmental protection, the findings of this work provide strong support for advancing more sustainable and efficient wastewater treatment technologies.

Overall, the successful application of the RFR model not only advances the practical use of ML technologies in wastewater treatment but also provides robust theoretical support for technological innovation and policy development in related fields. With continuous improvement and optimization of these predictive models, more efficient, cost-effective, and environmentally friendly wastewater treatment solutions are expected, contributing to global water resource management and sustainable development.

While the models presented perform excellently, they still have some limitations, mainly related to the dataset's constraints, model applicability, seasonal factors, and costs. First, the sample size and scope of the dataset are limited and may not fully represent the situation in all wastewater treatment plants. Additionally, noise and missing values in the data may affect the stability and accuracy of the models. Future improvements could involve collecting more diverse data to enhance the models. Besides, while the RFR model performs well, it may perform differently under varying wastewater types or treatment processes. Therefore, adjustments to the model or exploration of alternative methods may be necessary depending on the specific context. Seasonal factors, such as temperature and precipitation, may also influence wastewater treatment effectiveness, which are not considered here. Future work could incorporate seasonal variables to enhance the model's adaptability. Lastly, wastewater treatment plants may face challenges related to data collection and computational resources when implementing these models. Therefore, balancing prediction accuracy and costs will be a challenge in the

future application of these models. In summary, while this work provides accurate predictive models, further consideration of these factors is needed in real-world applications to achieve better outcomes.

Conclusion

This work analyzes the key water quality parameters in the urban wastewater treatment process of Northwest China and evaluates the performance of different regression models in predicting effluent water quality. The prediction performance for COD, NH_4^+-N, TP, and TN indicators is specifically compared in detail. The results indicate that the RFR model performs excellently in predicting various water quality parameters, demonstrating high prediction accuracy and stability. Specifically, for COD prediction, the RFR model achieves an MAE of 0.00934, MSE of 0.00198, and an R^2 value of 0.99954. For NH_4^+-N, TP, and TN indicators, the RFR model's MAE values are 6.76E−4, 0.000357, and 0.00251, respectively; the MSE values are 1.896E-6, 5.4E−7, and 4.1496E−5, respectively; and the R^2 values reach 0.99989, 0.99988, and 0.99989, respectively. When compared to five other common regression models, the RFR model consistently demonstrates superior performance. Especially in COD prediction, its MAE and MSE are significantly lower than those of the other models, and its prediction accuracy for NH_4^+-N, TP, and TN is also notably better.

Additionally, based on the mass balance method, this work quantitatively assesses the impact of the new method on carbon emissions in wastewater treatment plants. The comparison between the existing method with the new method reveals that the new approach significantly reduces carbon emissions in the wastewater treatment process, with an overall reduction of 17.33%. This finding provides an important reference for wastewater treatment plants adopting low-carbon emission technologies. It demonstrates that the new method can effectively reduce the carbon footprint while improving treatment efficiency, thus contributing to wastewater management and environmental protection. Overall, the main contribution of this work lies in the introduction of high-precision predictive models and carbon emission analysis, providing practical and quantifiable guidance for optimization and policymaking in the wastewater treatment sector. In the future, as datasets and models are further refined, the RFR model is expected to become an essential tool in wastewater management, helping decision-makers achieve better carbon reduction goals while handling wastewater. Furthermore, the findings also offer valuable insights for the development of environmental protection technologies in similar fields, and advance technological progress and green development in the environmental protection industry.

Data availability

The data presented in this study are available on request from the corresponding author.

Received: 19 September 2024; Accepted: 12 December 2024
Published online: 28 December 2024

References

1. Ferrentino, R., Langone, M., Fiori, L. & Andreottola, G. Full-scale sewage sludge reduction technologies: a review with a focus on energy consumption. *Water* **15**, 615. https://doi.org/10.3390/w15040615 (2023).
2. Zhou, C., Yu, Z. & Wang, Q. Analysis of temporal and spatial changes and influencing factors of sewage treatment rates of small towns in chongqing. *Front. Environ. Sci.* **11**, 1066371. https://doi.org/10.3389/fenvs.2023.1066371 (2023).
3. Patil, S. et al. Characterization and removal of microplastics in a sewage treatment plant from Urban Nagpur, India. *Environ. Monit. Assess.* **195**, 47. https://doi.org/10.1007/s10661-022-10680-x (2022).
4. Cheng, Q., Chunhong, Z. & Qianglin, L. Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor. *Sci. Rep.* **13**, 9149. https://doi.org/10.1038/s41598-023-36333-8 (2023).
5. Zhang, Q. et al. Dynamic decision-making for inspecting the quality of treated sewage. *Urban Clim.* **53**, 101752. https://doi.org/10.1016/j.uclim.2023.101752 (2024).
6. Bai, M., Li, W. & Xu, J. Research on greenhouse gas emission reduction methods of SBR and anoxic oxic urban sewage treatment system. *Sustainability* **15**, 7234. https://doi.org/10.3390/su15097234 (2023).
7. Dui, H., Zhu, Y. & Tao, J. Multi-phased resilience methodology of urban sewage treatment network based on the phase and node recovery importance in IoT. *Reliab. Eng. Syst. Saf.* **247**, 110130. https://doi.org/10.1016/j.ress.2024.110130 (2024).
8. Marin, E. & Rusănescu, C. O. Agricultural use of urban sewage sludge from the wastewater station in the municipality of Alexandria in Romania. *Water* **15**, 458. https://doi.org/10.3390/w15030458 (2023).
9. Su, Q. et al. Water–energy–carbon nexus: greenhouse gas emissions from integrated urban drainage systems in China. *Environ. Sci. Technol.* **57**, 2093–2104. https://doi.org/10.1021/acs.est.2c08583 (2023).
10. Xian, C., Gong, C., Lu, F., Wu, H. & Ouyang, Z. The evaluation of greenhouse gas emissions from sewage treatment with urbanization: understanding the opportunities and challenges for climate change mitigation in China's Low-carbon Pilot City, Shenzhen. *Sci. Total Env.* **855**, 158629. https://doi.org/10.1016/j.scitotenv.2022.158629 (2023).
11. Jiménez-Benítez, A. et al. A semi-industrial AnMBR plant for urban wastewater treatment at ambient temperature: analysis of the filtration process, energy balance and quantification of ghg emissions. *J. Environ. Chem. Eng.* **11**, 109454. https://doi.org/10.1016/j.jece.2023.109454 (2023).
12. Mahanna, H. et al. Prediction of wastewater treatment plant performance through machine learning techniques. *Desalin. Water Treatment* **319**, 100524. https://doi.org/10.1016/j.dwt.2024.100524 (2024).
13. Zhang, Y. et al. Machine learning modeling for the prediction of phosphorus and nitrogen removal efficiency and screening of crucial microorganisms in wastewater treatment plants. *Sci. Total Env.* **907**, 167730. https://doi.org/10.1016/j.scitotenv.2023.167730 (2024).
14. Cechinel, M. A. P. et al. Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: a plant-level analysis. *J. Water Process Eng.* **58**, 104758. https://doi.org/10.1016/j.jwpe.2023.104758 (2024).
15. Rios-Fuck, J. V. et al. Predicting effluent quality parameters for wastewater treatment plant: a machine learning-based methodology. *Chemosphere* **352**, 141472. https://doi.org/10.1016/j.chemosphere.2024.141472 (2024).
16. Shao, S. et al. Analysis of machine learning models for wastewater treatment plant sludge output prediction. *Sustainability* **15**, 13380. https://doi.org/10.3390/su151813380 (2023).
17. Chen, S., Yu, L., Zhang, C., Wu, Y. & Li, T. Environmental impact assessment of multi-source solid waste based on a life cycle assessment, principal component analysis, and Random Forest Algorithm. *J. Environ. Manage.* **339**, 117942. https://doi.org/10.1016/j.jenvman.2023.117942 (2023).

18. Wang, D. et al. Successful prediction for coagulant dosage and effluent turbidity of a coagulation process in a drinking water treatment plant based on the elman neural network and random forest models. *Environ. Sci. Water Res. Technol.* **9**, 2263–2274. https://doi.org/10.1039/D3EW00181D (2023).
19. Wu, X. et al. Coupling process-based modeling with machine learning for long-term simulation of wastewater treatment plant operations. *J. Environ. Manage.* **341**, 118116. https://doi.org/10.1016/j.jenvman.2023.118116 (2023).
20. Aghdam, E. et al. Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques. *J. Clean. Prod.* **405**, 137019. https://doi.org/10.1016/j.jclepro.2023.137019 (2023).
21. Sakti, A. D. et al. Optimizing city-level centralized wastewater management system using machine learning and spatial network analysis. *Environ. Technol. Innov.* **32**, 103360. https://doi.org/10.1016/j.eti.2023.103360 (2023).
22. Piłat-Rożek, M. et al. Application of machine learning methods for an analysis of e-nose multidimensional signals in wastewater treatment. *Sensors* **23**, 487. https://doi.org/10.3390/s23010487 (2023).
23. Bellamoli, F., Di Iorio, M., Vian, M. & Melgani, F. Machine learning methods for anomaly classification in wastewater treatment plants. *J. Environ. Manage.* **344**, 11859. https://doi.org/10.1016/j.jenvman.2023.118594 (2023).
24. Chauhan, J. et al. Gradient-boosted decision tree with used slime mould algorithm (sma) for wastewater treatment systems. *Water Reuse* **13**, 393–410. https://doi.org/10.2166/wrd.2023.046 (2023).

Author contributions

JS: Writing—review & editing, Conceptualization, Methodology; XS: Writing—review & editing, Conceptualization, Methodology, Funding acquisition, Resources; XG: Writing—review & editing, Data curation, Formal analysis, Project administration; XC: Writing —original draft, Data curation, Formal Analysis; YT: Writing—original draft, Visualization; JL: Writing —original draft, Validation. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by National Natural Science Foundation of China on "A Study on Social Capital, CEO Power, Corporate Resource Allocation & Its Economic Consequences" (Grant No. 20CGL011), and the Ministry of Education Industry-Academia Cooperation and Collaborative Talent Cultivation Project (Grant No. 231101339153047), and the Chinese Postdoctoral Science Fund 74th (Grant No. 2023M740792).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.S.

Reprints and permissions information is available at www.nature.com/reprints.

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