Heliyon 10 (2024) e33695

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Assessing and predicting water quality index with key water parameters by machine learning models in coastal cities, China

Jing Xu [a,*], Yuming Mo [b], Senlin Zhu [a], Jinran Wu [c], Guangqiu Jin [d], You-Gan Wang [e], Qingfeng Ji [a], Ling Li [f]

[a] College of Hydraulic Science and Engineering, Yangzhou University, Yangzhou, China
[b] School of Naval Architecture and Ocean Engineering, Jiangsu University of Science and Technology, Zhenjiang, China
[c] Institute for Positive Psychology and Education, Australian Catholic University, North Sydney, Australia
[d] The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing, China
[e] School of Mathematics and Physics, The University of Queensland, Queensland, Australia
[f] Key Laboratory of Coastal Environment and Resources of Zhejiang Province (KLaCER), School of Engineering, Westlake University, Hangzhou, China

ARTICLE INFO

Keywords:
Water quality
Key water parameters
Water quality index (WQI)
Machine learning models
Coastal cities

ABSTRACT

The water quality index (WQI) is a widely used tool for comprehensive assessment of river environments. However, its calculation involves numerous water quality parameters, making sample collection and laboratory analysis time-consuming and costly. This study aimed to identify key water parameters and the most reliable prediction models that could provide maximum accuracy using minimal indicators. Water quality from 2020 to 2023 were collected including nine biophysical and chemical indicators in seventeen rivers in Yancheng and Nantong, two coastal cities in Jiangsu Province, China, adjacent to the Yellow Sea. Linear regression and seven machine learning models (Artificial Neural Network (ANN), Self-Organizing Maps (SOM), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB) and Stochastic Gradient Boosting (SGB)) were developed to predict WQI using different groups of input variables based on correlation analysis. The results indicated that water quality improved from 2020 to 2022 but deteriorated in 2023, with inland stations exhibiting better conditions than coastal ones, particularly in terms of turbidity and nutrients. The water environment was comparatively better in Nantong than in Yancheng, with mean WQI values of approximately 55.3–72.0 and 56.4–67.3, respectively. The classifications "Good" and "Medium" accounted for 80 % of the records, with no instances of "Excellent" and 2 % classified as "Bad". The performance of all prediction models, except for SOM, improved with the addition of input variables, achieving R^2 values higher than 0.99 in models such as SVM, RF, XGB, and SGB. The most reliable models were RF and XGB with key parameters of total phosphorus (TP), ammonia nitrogen (AN), and dissolved oxygen (DO) ($R^2 = 0.98$ and 0.91 for training and testing phase) for predicting WQI values, and RF using TP and AN (accuracy higher than 85 %) for WQI grades. The prediction accuracy for "Medium" and "Low" water quality grades was highest at 90 %, followed by the "Good" level at 70 %. The model results could contribute to efficient water quality evaluation by identifying key water parameters and facilitating effective water quality management in river basins.

* Corresponding author.
E-mail address: xujing7503@yzu.edu.cn (J. Xu).

https://doi.org/10.1016/j.heliyon.2024.e33695
Received 21 January 2024; Received in revised form 14 June 2024; Accepted 25 June 2024
Available online 27 June 2024

1. Introduction

With the rapid urbanization, industrialization and agricultural activities, large amount of river pollution from point or non-point sources have posed increasingly challenges over the world [1,2]. The ongoing deterioration of water quality has put safe water supplies at risk, causing water pollution incidents and damaging aquatic ecosystems [3,4], especially in coastal cities with more prosperous economic development and intensive anthropogenic activities [5,6].

Researchers have reported water pollution problems in various coastal cities worldwide. For instance, discharge of domestic sewage and industrial effluents in Surat, which locates in river Tapi and extends up to the Arabian Sea in western India [7]; Sevastopol Bay near the Black Sea has high levels of petroleum hydrocarbons and organic matter [8]; while Boston Harbor in the USA suffers from nutrient (N and P) and organic pollution [9]. Similarly, water pollution exists in coastal cities in China due to earlier development and stronger policy support compared to inland cities [10]. For example, Wu et al. [11] used cluster analysis and fuzzy logic approach to evaluate trophic status in Daya Bay, South China Sea, indicated that Chlorophyll *a* and phosphorus were major water pollution; Liu et al. [12] applied Organic Pollution Index and Eutrophication Index to quantify water quality status and reported heavy water pollution near Luanhe River in Hebei Province, western Bohai Sea; Zhang et al. [13] reviewed heavy metal pollution in coastal cities of the East China Sea and found that Cd pollution was serious and posed ecological risks in Hangzhou and Quanzhou. The water quality deterioration problems in the Yellow Sea have also garnered attention [14]. Dong et al. [15] analyzed a 16-month water quality dataset in Yantai, Shandong Province and reported that water temperature and nutrients were main causes of trophic status in coastal waters; Sun et al. [16] evaluated spatiotemporal variations of nutrients in the northern Yellow Sea (Shandong Province) and highlighted the influence of human activities. While previous studies have primarily focused on water pollution in Shandong Province adjacent to the Yellow Sea [17–19], few have comprehensively evaluated water quality in coastal cities in Jiangsu Province, which has millions of residents and is an important part of China's coastal economic belt [20], suffering from serious water deterioration problems [21].

A scientific assessment of water quality is fundamental for local governments to take effective and efficient measures in river management [22,23]. Recently, evaluation methods have been developed from separate water quality parameters to comprehensive indices, such as pollution index (PI), heavy metal pollution index (HPI), fuzzy comprehensive evaluation, best management practices (BMPs), total maximum daily loads (TMDLs) and the European Water Framework Directive [24–28]. Among these, the Water Quality Index (WQI) is a non-dimensional index derived from various water quality parameters such as water temperature, pH, dissolved oxygen, total suspended solids, ammonia nitrogen and total phosphorus, depending on the availability of data, has become one of the most popular tools for evaluating water quality due to its simple architecture compared to physical-processed or hydrological models [29–31]. The first WQI model was developed in the 1960s using ten water quality parameters and was later modified to a more rigorous version defining parameter selection and weighting [32,33]. As WQI models were applied in different areas and fields, several revisions were made based on the original, such as BCWQI developed by the British Columbia Ministry for Environment, Lands and Parks or CCME WQI developed by the Canadian Council of Ministers of the Environment [34,35]. Based on summary from Uddin et al. [36], WQI calculation process involves selecting water quality parameters, determining parameter weights, and computing the overall index. Many researchers employed WQI tools to assess water quality in rivers or lakes in China, such as Lake Taihu, Lake Chaohu, Dongjiang River and Luanhe River [37–40]. These studies involved selecting water quality parameters and modifying the corresponding relative weights and normalization factors to suit the conditions of water bodies in China, based on previous literature [41,42].

While the WQI provides an easy-to-understand statistical information and comprehensive evaluation of water quality status by integrating biophysical and chemical indicators [43,44], it requires as many water quality parameters as possible, which necessitates considerable time and cost for field sample collection and laboratory analysis [39,45]. To address this limitation, the WQI_{min} was proposed, where key water parameters are selected from the original set to calculate WQI_{min}, which had applied by many researchers [42,46]. For example, Chen et al. [47] selected six key parameters from ten water quality parameters by principal component analysis to calculate the WQI_{min} in Huaihe River, Northeast China; Pan et al. [48] and Qi et al. [49] applied stepwise regression to selected six water parameters from fifteen and five from ten, with predicted R^2 of 0.938 and 0.903 respectively. While previous studies have highlighted the benefits of time and cost savings by predicting the WQI using five or six key water parameters selected through traditional statistical analysis methods, there are situations where it may be beneficial to further reduce the number of key parameters selected. For instance, the ability to quickly evaluate water quality status with minimal field or laboratory measurements is crucial during sudden water pollution incidents or when there is a need to alert authorities about deteriorating water quality [50].

Over the decades, with the development of state-of-the-art artificial intelligence algorithms, machine learning models have been widely applied in water resources management due to their advantages in describing non-linear or complex relationships between input and output variables without considering physical mechanism or dynamic processes [51–53]. Many studies have developed machine learning models for water quality prediction, such as Kulisz and Kujawska [54] used five physicochemical parameters as input layers to predict WQI in river Warta, Poland by artificial neural networks (ANN) and Najafzadeh and Ghaemi [55] applied support vector machines (SVM) to predict biochemical/chemical oxygen demand by nine parameters, highlighting the successes of artificial intelligence techniques. Recently, some studies have recommended that ensemble tree-based algorithms, such as random forest (RF) and extreme gradient boosting (XGB) performed better in WQI prediction [56,57]. The great potential of machine learning models makes it possible to apply even fewer key parameters to obtain higher precision in WQI prediction compared to the WQI_{min} model [58,59]. The factors affecting model performance are not only related to the model type or dataset size but also depend on the input independent variables used in developing the machine learning models, highlighting the importance of comparing model performances between different input datasets [50]. However, to our knowledge, few studies have focused on developing machine learning models to predict the WQI and comparing these performances with large dataset in coastal cities adjacent to the Yellow Sea, especially

in Jiangsu Province, one of the most developed regions in China suffering from water pollution problems [60,61]. These achievements are crucial in determining the minimum numbers of key water parameters and the most reliable model to achieve the highest precision in prediction of WQI values or grades (such as "excellent", "Good" or "Bad" water quality classified by different ranges of WQI values), which could provide a time- and cost-saving way in efficient and effective water environment protection.

In the study, a four-year water quality dataset (2020–2023) containing nine water quality indicators from seventeen rivers in Yancheng and Nantong, two coastal cities in Jiangsu Province adjacent to the Yellow Sea, was collected. The main objectives were to (1) develop statistical and machine learning models to predict the WQI based on various groups of input variables; (2) select the key water parameters and most reliable models in predicting WQI values and grades; and (3) assess spatio-temporal variations of key water parameters in the two coastal cities. The research is expected to contribute to effective and efficient river management in the coastal cities adjacent to the Yellow Sea, and the methodological framework can also be applied in other rivers worldwide.

2. Material and methods

2.1. Study area

Yancheng and Nantong are two coastal cities in Jiangsu Province, east-central China, situated east of the Yellow Sea (Fig. 1). Yancheng (119.45–120.90°N, 32.57–34.47°E) is located in the middle of the northern plain of Jiangsu Province, covering an area of 16,931 km^2 with over six million inhabitants [62]. Yancheng has the largest land area and longest coastline in Jiangsu Province, occupying 70 % of the tidal flat and 56 % of the coastline length [63]. The terrain of Yancheng is plain landform, including the Huanghuai Plain, the Lixiahe Plain, and the Binhai Plain. Over 1000 m of loose alluvial deposits, consisting of clay, sub-clay, and sandstone, were formed in Yancheng due to the rise and fall of sea levels and the interaction of the Yangtze River, Yellow River, and Huai River throughout history [64].

Nantong is situated south of Yancheng, at the forefront of the Yangtze River Delta alluvial plain, lying between 31.68–32.77°E and 120.20–121.91°N, with an area of 8001 km^2 and a population of over eight million [65]. The average altitude of Nantong city is below 5 m, with vast marine plains, salt fields and beaches. Yancheng is in the north-to-south transitional zone from a temperate to sub-tropical monsoon climate, while Nantong lies in the subtropical monsoon climate zone. The average temperature ranges from 13.7 to 15.1 °C, with an annual precipitation of around 1000 mm and an average evaporation capacity of approximately 1300 mm [66,67].

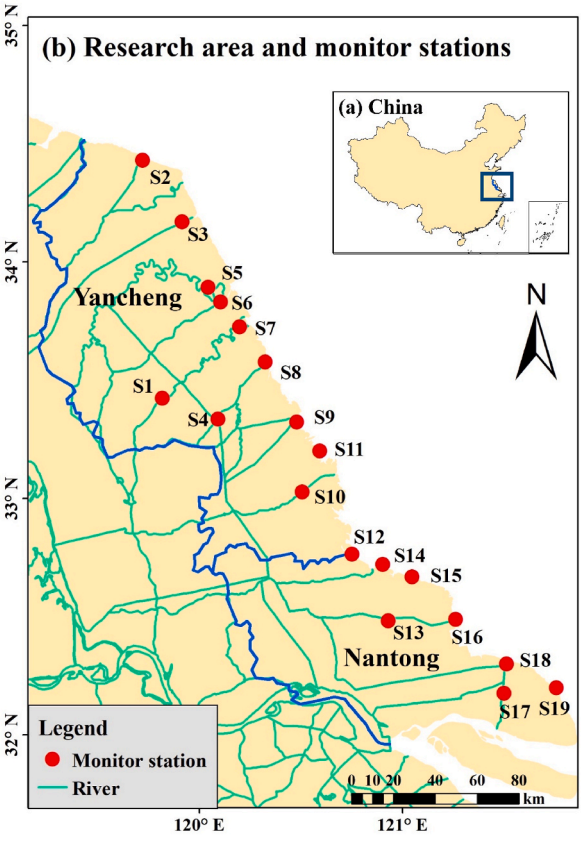

Fig. 1. The map of research area which is lying in eastern China and the locations of nineteen water quality monitor stations in Yancheng and Nantong cities.

Both cities have numerous rivers, such as Mangshe River and Sheyang River in Yancheng, and Beiling River and Tongqi Canal in Nantong, with high river network density of 3.1 km/km^2 and 2.22 km/km^2, respectively, mainly flowing into the Yellow Sea [68]. The Sheyang River is one of the largest rivers in Jiangsu Province, with a maximum depth of about 8 m [69]. The thickness of the aquifer is about 10–15 m and depth of groundwater is about 1–4 m in Yancheng [70]. The groundwater depth in Nantong is about 0.45–2.88 m, influenced by annual periodic changes of rainfall and tidal water levels along the Yangtze River [71].

Farmland is the dominated land use type in Yancheng, with paddy fields and dry land accounting for approximately 35 % and 43 % of the area, respectively [72]. Farmland accounts for a lesser percentage in Nantong, around 44 % [73]. In addition to non-point pollution caused by agricultural cultivation, rapid urbanization and economic development in the two cities recently caused serious problems in environmental deterioration and ecological damage. Simultaneously, several water pollution incidents have occurred, posing risks to safe drinking water sources for more than ten million people [74,75].

2.2. Water quality dataset and WQI calculation

To evaluate the water environment in rivers and select the key water parameters for predicting the WQI in the two coastal cities, water quality data from nineteen monitor stations (S1 to S19) were collected from Environmental Protection Department in Jiangsu province (see Table 1). Stations S1 to S11 are located in Yancheng, while S12 to S19 are in Nantong. In Yancheng, eight monitor stations (S2, S3, S5, S6, S7, S8, S9 and S11) were situated at coastal ports alongside the Yellow Sea, with S1 and S4 being upstream water quality stations for S7 and S8, respectively. In Nantong, two costal port stations (S18 and S19), one port station to the Yangtze River (S13), three stations at floodgates in the coastal regions (S12, S15 and S16) and two stations at canals flowing into sea (S14 and S17) were included.

Nine water quality parameters were considered in the research based on data availability and literature on WQI calculation: water temperature (WT/($^\circ C$)), pH, dissolved oxygen (DO/(mg/L)), electrical conductivity (EC/(μ s/cm)), turbidity (Tur/(NTU)), permanganate index (COM/(mg/L)), ammonia nitrogen (AN/(mg/L)), total phosphorus (TP/(mg/L)) and total nitrogen (TN/(mg/L)) [76,77]. The dataset ranged from November 2020 to August 2023 and included a total of 18084 records. To compare the water quality status in Yancheng and Nantong, a *t*-test was used to analyze whether significant differences existed between the two cities.

The WQI values were calculated by:

$$WQI = \frac{\sum_{i=1}^{n} C_i P_i}{\sum_{i=1}^{n} P_i} \tag{1}$$

where the C_i represents normalized value of parameter *i* and P_i represents the weight of parameter *i*. The normalized values and weights of all water quality parameters are shown in Table S1, which were recommended by previous literature evaluating water quality in the area having similar meteorology and hydrology conditions with the research area in the article [37,38]. Based on the calculated WQI values, the water quality can be categorized into five grades (Table S2), accordingly, "excellent" level (91–100), "good" level (71–90), "medium" level (51–70), "low" level (26–50) and "bad" level (0-25).

2.3. Model development

In the research, various models were applied to predict the WQI using different input groups of water quality parameters, including Linear regression Model (LM) and machine learning models. The LM is a simple and fundamental model in statistical analysis, widely used to describe linear relationships between one or more independent variable and dependent variable. Seven common machine learning models were applied for WQI prediction: Artificial Neural Network (ANN), Self-Organizing Maps (SOM), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB), Stochastic Gradient Boosting (SGB) [78,79].

ANN is a popular machine learning models reported to be an efficient and accurate tool for water quality analysis due to its robustness and capability in solving nonlinear and nonstationary problems [80,81]. An ANN model consists of three kinds of layers (input, hidden, and output), which can increase the ability to describe complex relationships. SVM is a kernel-based model that includes regularization and kernel function (linear, polynomial or radial basis function), providing high flexibility and global optimization by constructing expert information [82,83]. KNN is a simple but computationally complex algorithm for classification and prediction, where nearby records tend to have similar values based on distance measures such as Euclidean, Manhattan, Minkowski distance [84]. SOM is a type of non-supervised artificial neural advantageous for visualizing high-dimensional data and interpreting nonlinear relationships between multi-dimensional data. In an SOM network, the input layer and output layer (competitive layer) are fully connected. In competition, the weights of the winning neuron and its neighboring neurons are adjusted following input patterns, allowing the model to achieve self-organizing learning capabilities [85,86]. RF is an ensemble learning algorithm consisting of multiple decision trees selected randomly from samples and features, contributing to improved prediction/classification accuracy and reduced model variance [86–88]. XGB and SGB are two types of ensemble tree methods that improve prediction accuracy by constructing multiple weak learners (classification and regression trees) using the gradient descent architecture [89,90].

To minimize the water quality parameters in WQI prediction, different groups of input variables, including partial or the entire water quality dataset, were collected. The input variables were determined by Pearson's correlation analysis, where the corresponding water quality parameters were added sequentially into input groups according to their correlation coefficients in descending order. These different input groups were set as independent variables to predict the WQI using various statistical and machine learning

models. The entire dataset was standardized (0–1 standardization) and then split into two group, with 70 % of the data used for model training and the remaining 30 % for model testing [91]. The developed models were validated by a 10-fold cross-validation approach, which is widely used for comparing and optimizing model performances during the development of machine learning models [92].

2.4. Model evaluation

The model performances and sensitivity were evaluated using mean absolute error (MAE), root mean squared error (RMSE) and coefficients of determination (R^2), which are common statistical measures employed in literatures [93–95]. Model efficiency was analyzed by the Nash–Sutcliffe efficiency (NSE) index, widely used in hydrological model assessment initially and recently utilized in selection the best machine learning models [96,97]. The calculation formulas are as follows [98]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - A_i| \tag{2}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(A_i - P_i)^2}{\sum_{i=1}^{n}(A_i - \overline{A_i})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{4}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n}(A_i - \overline{A_i})(P_i - \overline{P_i})\right]^2}{\sum_{i=1}^{n}(A_i - \overline{A_i})^2 \sum_{i=1}^{n}(P_i - \overline{P_i})^2} \tag{5}$$

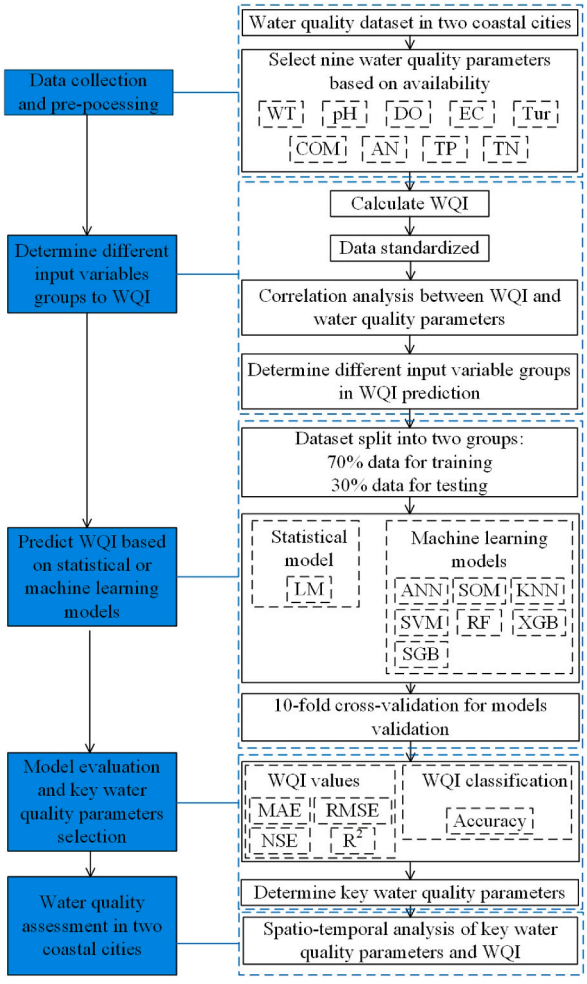

Fig. 2. The methodological framework for WQI prediction and key water quality parameters selection in the two coastal cities.

where A_i and P_i represent the actual and predicted WQI at the ith data; $\overline{A_i}$ and $\overline{P_i}$ represent the average value of the actual and predicted WQI respectively; and n represents the number of WQI records.

To analyze the levels of model uncertainty, four statistical indices, including average, maximum, minimum and standard deviation, were calculated at all stations [99,100]. Overall, five steps were considered in the research: data collection and pre-processing, determining different input variables groups for WQI prediction, development of WQI prediction models, selection of key water parameters and the most reliable model, and assessment of key water parameters in the two coastal cities. The methodological framework of the research is shown in Fig. 2. In the study, all data processing and model developments were completed using the R statistical computing platform [101].

3. Results and discussion

3.1. Summary of water quality indicators

The summary of nine water quality parameters and corresponding WQI at nineteen monitor stations were shown in Table 2. The average WT ranged from 17.6 °C to 19.9 °C with the highest WT recorded in Nantong (S12 and S18) and the lowest WT in Yancheng (S3, S5 and S9). The WT in Yancheng was significantly lower than in Nantong (Table 3), mainly due to the geographic locations of the cities (Yancheng is north to Nantong city). The S1 and S4 had the lowest pH (7.51), while relatively higher pH existed at S2, S13 and S19. The dissolved oxygen (DO) concentration showed no significant difference between the two cities, but variations existed among the monitoring stations. Accordingly, the S10 experienced the worst DO conditions (5.93 mg/L), while S3 had the best DO conditions (8.71 mg/L), implying the presence of biodegradable organic pollution in the Dongtai River [102].

The EC tended to be higher at stations closer to the sea. Consequently, S1, S4 and S13, which were upstream stations of estuary of the Yellow Sea or ports to the Yangtze River, had relatively lower EC (<700 μS/cm), while the ports or floodgates in the coastal regions had higher EC (ranging from 1330 to 2263 μS/cm). Turbidity values were significant higher in Yancheng than in Nantong, with S7 in Yancheng having the highest turbidity (91.8 NTU) and S17 in Nantong having the lowest (12.3 NTU). Turbidity, an index of light scattering by suspended particles, has been seen as a surrogate index for suspended sediment in waterbody, and the high turbidity was an important cause in formation of coastal tidal flat wetlands in Yancheng [103,104]. Additionally, the construction of floodgates in Xinyang Port altered the hydrodynamic characteristics of the river and resulted to sediment deposition alongside the river section, providing an explanation for the highest turbidity at S7 [105]. COM and TN concentrations in Yancheng were significantly higher than in Nantong, while AN and TP exhibited the opposite pattern.

According to China's Environmental Quality Standards for Surface Water (GB 3838-2002), water quality at most stations failed to meet the standard of V grade because of high TN concentration. Due to river nitrogen exports were the largest pollution sources in the Yellow Sea, our result indicated the importance to control TN pollution in the two coastal cities in order to protect the costal water ecosystem [106].

3.2. WQI characteristics

3.2.1. WQI values and grades in the two coastal cities

The WQI values were slightly higher in Nantong (mean values about 55.3–72.0) than in Yancheng (mean values about 56.4–67.3), implying that the water environment was relatively better in Nantong. Based on the WQI values at all monitor stations, the grades of

Table 1
Details of water quality monitoring stations in the research area.

Station name	Code	Longitude	Latitude	River	River length (km)	Area(km^2)
Fenghuangqiao	S1	120.048	33.359	Mangshe River	46	870
Touzengzha	S2	120.092	34.373	Abandoned Yellow River	728	4291
Liuduozha	S3	120.255	34.093	Main Irrigation Channel of North Jiangsu	168	564
Datuanqiao	S4	120.316	33.243	Doulong Port	4428	4428
Sheyanghezha	S5	120.346	33.804	Sheyang River	198	4036
Huangshagangzha	S6	120.399	33.736	Huangsha Port	89	865
Xinyanggangzha	S7	120.481	33.621	Xinyang Port	70	2478
Doulonggangzha	S8	120.588	33.460	Doulong Port	55	4428
Wanggangzha	S9	120.708	33.190	Wanggang River	44	498
Fuminqiao	S10	120.774	32.690	Dongtai River	55	479
Chuandongzha	S11	120.805	33.057	Chuandong Port	49	648
Beilingxinzha	S12	120.953	32.606	Beiling River	45	323
Jiuweigangqiao	S13	121.042	32.306	Jiuwei Port	47	2123
Xiaoyangkou	S14	121.049	32.567	Bencha Canal	79	446
Huandongzhakou	S15	121.186	32.479	Jucha River	17	300
Donganzhaqiao	S16	121.376	32.278	Rutai Canal	155	452
Junandaqiao	S17	121.568	31.941	Tongqi Canal	93	530
Dayanggangqiao	S18	121.598	32.063	Tonglv Canal	79	2294
Tanglugangzha	S19	121.829	31.936	Tongqi Canal	93	530

Table 2
Summary of nine water quality parameters and WQI at each monitoring station.

Code	WT (°C)	pH	DO (mg/L)	EC (μS/cm)	Tur (NTU)	COM (mg/L)	AN (mg/L)	TP (mg/L)	TN (mg/L)	WQI	Records number
S1	18.2 ± 8.4	7.51 ± 0.31	6.91± 3.20	572 ± 83	41.3 ± 18.0	4.25 ± 1.17	0.19 ± 0.14	0.10 ± 0.03	2.00 ± 0.53	61.4 ± 8.9	976
S2	17.8 ± 8.7	8.06 ± 0.39	8.49± 3.09	652 ± 126	59.0 ± 36.9	3.73 ±0.71	0.08 ± 0.06	0.07 ± 0.03	1.81 ± 0.56	67.0 ± 7.6	897
S3	17.5 ± 8.7	7.95 ±0.42	8.71 ± 2.86	659 ± 154	79.9 ± 69.0	3.66 ± 1.17	0.10 ± 0.08	0.08 ± 0.05	1.77 ± 0.57	67.3 ± 6.9	961
S4	18.4 ± 8.4	7.51 ±0.31	6.60 ±2.79	664 ± 129	58.8 ± 20.1	4.18 ± 1.39	0.22 ±0.22	0.14 ±0.06	2.13 ± 0.58	58.9 ± 8.0	982
S5	17.6 ± 8.4	7.78 ±0.35	7.62 ±3.10	888 ±386	55.0 ± 35.9	5.55 ±1.41	0.28 ±0.23	0.13 ± 0.05	2.56 ±0.58	58.1 ±10.1	954
S6	17.9 ± 8.7	7.73 ± 0.41	6.62 ±3.39	921 ± 252	39.5 ± 37.1	5.41 ± 1.40	0.28 ±0.26	0.17 ± 0.06	2.49 ± 0.52	56.4 ±10.5	957
S7	17.9 ± 8.4	7.74± 0.36	7.75 ± 3.20	1032 ± 446	91.8 ± 43.8	4.94 ± 1.16	0.23 ±0.17	0.13 ± 0.05	2.62 ± 0.63	56.6 ± 7.8	951
S8	18.4 ± 8.6	7.78 ± 0.41	6.92 ± 3.25	1024 ± 462	41.4 ± 20.0	4.62 ± 1.57	0.25 ±0.27	0.13 ± 0.06	2.53 ± 0.69	58.9 ±10.9	958
S9	17.6 ± 8.4	7.85 ± 0.41	7.38 ± 4.11	1330 ± 518	27.4 ± 16.0	5.21 ±1.39	0.32 ±0.34	0.17 ± 0.09	2.91 ± 1.04	57.7 ±11.6	911
S10	18.7 ± 8.7	7.75 ± 0.42	5.93 ± 2.64	904 ± 383	29.3 ± 37.2	4.23 ±1.47	0.15 ±0.23	0.14 ± 0.10	2.21 ± 0.70	62.4 ±10.7	942
S11	18.9 ± 8.5	7.9 2 ± 0.47	8.18 ±3.17	1532 ± 809	37.9 ± 31.2	5.06 ± 1.53	0.31 ±0.40	0.17 ± 0.11	3.29 ± 1.04	59.6 ±11.1	927
S12	19.9 ± 7.9	7.91 ±0.39	7.89 ±4.20	1675 ±496	23.5 ±15.0	6.47 ±1.25	0.54 ±0.55	0.26 ±0.15	2.82 ±1.06	55.3 ±10.9	926
S13	19.6 ± 7.2	8.08 ±0.17	7.74 ±1.53	361 ±56	47.3 ±68.6	1.81 ±0.59	0.07 ±0.14	0.09 ±0.04	2.23 ±1.97	72.0 ±6.3	974
S14	19.3 ± 7.8	7.82 ±0.36	6.97 ±3.84	1100 ±282	27.9 ±19.8	5.39 ±1.38	0.45 ±0.54	0.21 ±0.12	2.50 ±1.83	57.2 ±11.5	948
S15	19.2 ± 8.1	7.86 ±0.41	7.62 ±4.79	1442 ±820	32.7 ±25.9	5.41 ±1.49	0.36 ±0.49	0.21 ±0.12	2.75 ±1.23	57.0 ±11.7	961
S16	18.9 ± 8.3	7.87 ±0.32	7.46 ±3.22	1066 ±732	37.3 ±28.9	4.28 ±1.48	0.36 ±0.39	0.19 ± 0.11	2.16 ±0.76	59.9 ±10.2	963
S17	19.1 ± 8.1	7.80 ±0.37	7.18 ±4.09	611 ±126	12.3 ±7.7	3.67 ±0.84	0.12 ±0.09	0.09 ±0.04	1.69 ±0.45	69.3 ±8.7	981
S18	19.9 ± 7.7	7.95 ±0.25	6.69 ±2.16	527 ±216	85.7 ±47.7	2.34 ±1.15	0.19 ±0.15	0.15 ±0.07	2.04 ±0.49	62.2 ±8.4	950
S19	18.7 ± 7.8	8.03 ±0.34	7.40 ±3.57	2263 ±1277	33.9 ±32.6	4.80 ±1.06	0.28 ±0.26	0.16 ±0.07	2.43 ±1.74	58.3 ±9.6	965

Table 3
The comparison of water quality in Yancheng and Nantong city with T-test.

Water quality parameter	Yanhceng	Nantong	p-value
WT (°C)	18.1	19.3	$<2 \times 10^{-16}$***
pH	7.78	7.92	$<2 \times 10^{-16}$***
DO (mg/L)	7.36	7.37	0.929
EC (μS/cm)	922	1127	$<2 \times 10^{-16}$***
Tur (NTU)	51.1	37.5	$<2 \times 10^{-16}$***
COM (mg/L)	4.62	4.26	$<2 \times 10^{-16}$***
AN (mg/L)	0.22	0.30	$<2 \times 10^{-16}$***
TP (mg/L)	0.13	0.17	$<2 \times 10^{-16}$***
TN (mg/L)	2.39	2.32	1×10^{-4}***
WQI	60.4	61.5	3×10^{-11}***

The asterisk means significant different existing between the two coastal cities.

WQI could be classified (Table S2). According to the average WQI values, the water quality grades were “Medium” at almost all stations except “Good” level in S12.

The percentage of each WQI grades in all stations were shown in Table 4. Based on the calculated WQI values, no record met the criteria of “Excellent” water quality. S13 had the most records of “Good” water quality, with percentage of 71.3 %, while S4 had the least, with a percentage of 0.9 %. The percentage of "Good" level water quality was below 5 % at four stations in Yancheng (S4, S5, S6, and S7) and one station in Nantong (S12). The “Medium” level occupied the most records, with the percentage exceeding 70 % at nine out of eleven monitoring stations (S1, S4, S5, S6, S7, S8, S9, S10, and S11) in Yancheng and three out of eight stations (S16, S18, and S19) in Nantong. The percentage of “Medium” water quality at other stations were less than 70 %, with a minimum was 27 % in the S13. Two stations (S6 and S9) in Yancheng and three stations in Nantong (S12, S14 and S15) had more than 20 % records assessed as “Low” level, while four stations (S2 and S3 in Yancheng and S13 and S17 in Nantong) had the least percentage, which less than 5 %. The percentage of “Bad” level were less than 2 % and even zero at all monitor stations.

3.2.2. Correlation between water quality parameters and WQI

To select the key water parameters that had an important effect on WQI prediction, the correlation between different water quality parameters and WQI values was analyzed (Fig. 3). TP was the variable with the most important negative influence on WQI, with a correlation coefficient of −0.72, followed by AN and DO, with correlation coefficients were −0.66 and 0.63, respectively. COM and pH were the next two parameters, with correlation coefficients of −0.56 and 0.54, respectively. Although TN was the most polluted contaminant in the research area based on the results in Section 3.1, it had a relatively weak impact on WQI values, with a correlation coefficient of −0.31. Tur and EC had the lowest correlation coefficients of −0.22 and −0.12, respectively, suggesting that they may have a minor influence on WQI predictions and could contribute to cost-saving in sample collection by reducing the sampling frequencies of these two parameters.

Based on the result of the correlation analysis, nine groups of input water quality parameters were determined, as shown in Table 5, where one variable was added to each group according to the descending order of correlation coefficients.

Table 4
summary of the WQI classification at all monitoring stations.

Station	Excellent (%)	Good (%)	Medium (%)	Low (%)	Bad (%)
S1	0.0	7.6	78.1	14.3	0.0
S2	0.0	38.4	59.1	2.6	0.0
S3	0.0	33.8	63.9	2.3	0.0
S4	0.0	0.9	87.3	11.7	0.1
S5	0.0	4.8	77.0	17.9	0.2
S6	0.0	2.0	72.1	25.7	0.2
S7	0.0	1.2	79.4	19.5	0.0
S8	0.0	9.9	70.1	19.8	0.1
S9	0.0	5.5	72.0	20.9	1.6
S10	0.0	16.7	71.0	11.9	0.4
S11	0.0	13.3	70.6	15.5	0.6
S12	0.0	1.6	68.7	28.9	0.8
S13	0.0	71.3	27.0	1.5	0.2
S14	0.0	6.5	69.3	23.6	0.5
S15	0.0	7.6	66.8	24.9	0.7
S16	0.0	9.9	74.6	15.5	0.1
S17	0.0	49.4	47.3	3.3	0.0
S18	0.0	8.7	82.1	9.2	0.0
S19	0.0	5.4	74.8	19.6	0.2

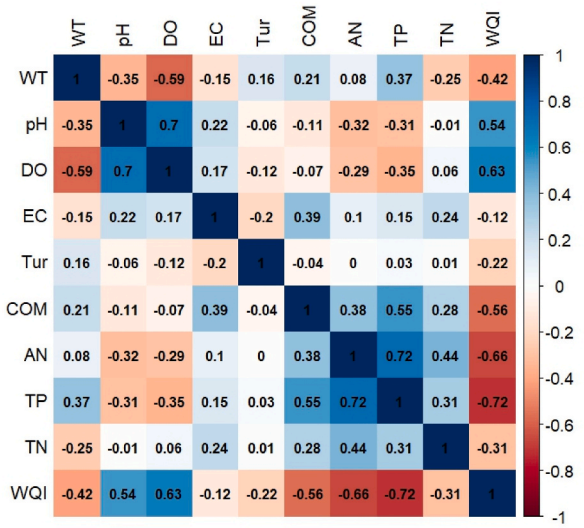

Fig. 3. The correlation coefficient between all water quality parameters and the WQI.

3.3. Model evaluation and selection of key water parameters

3.3.1. Model evaluation in prediction of WQI values

The nine input groups, which included different numbers of water quality parameters, acted as independent variables for LM and seven machine learning models. All models were validated by 10-fold cross-validation, and the optimized tuning parameters were summarized in Table 6. The performance metrics, including MAE, NSE, RMSE and R^2 from the training and testing phases, were shown in Table S3.

For the LM, the MAE and RMSE decreased with an increasing number of input variables, while NSE and R^2 increased from 0.513 to 0.825 in the training phase and from 0.508 to 0.838 in the testing phase, indicating that model performance improved with the addition of input variables. The performances of most machine learning models were similar to the LM, except for the SOM, whose performance improved with an increase in input variables from one to four (TP, AN, DO and COM) but declined when more variables were added. This may be related to model optimization to different local optima, resulting in diverse model performance caused by various input groups [107,108]. The R^2 and NSE for all models that predicted using one variable (TP) ranged from 0.5 to 0.6, except for KNN, where the NSE values were 0.412 and 0.381 for the training and testing phases, respectively. For LM and SOM, the highest values of NSE and R^2 were about 0.8 with nine input variables (all water quality parameters) and four variables (TP, AN, DO, COM) respectively. The NSE and R^2 were higher than 0.95 for the other models applying nine water quality parameters to predict the WQI, and the NSE and R^2 from RF, XGB and SGB even higher than 0.99. Similar WQI prediction with ten water quality parameters in North of Vietnam reported R^2 from RF and SVM were 0.92 and 0.72 respectively, however, it pointed out that RF seemed to have the best model performance [109]. The application of RF models in prediction of river WQI were also indicated by previous studies, such as six water quality variables to predict WQI grades in the Klang River Basin, Malaysia [110] and prediction from twelve input combinations of water quality parameters (the highest $R^2 = 0.998$) in the Illizi region, Algerian [111].

The MAE ranged from 0.07 to 0.1 for one input variable and decreased with the addition of water quality parameters, reaching less than 0.05 or even 0.01 (XGB) with nine input variables except SOM. The RMSE values had similar characteristics to the MAE and the lowest value was obtained with nine water quality parameters by XGB (0.006 in train phase and 0.011 in testing phase). Khoi et al. [112] highlighted the highest accuracy of WQI prediction by XGB ($R^2 = 0.989$ and RMSE = 0.107) among boosting-based, decision tree-based and ANN-based algorithms in La Buong River. Analysis of water quality in Ujjain city, India reported that XGB had

Table 5
Different combinations of input variables to predict the WQI.

Group number	Input variables
1	TP
2	TP, AN
3	TP, AN, DO
4	TP, AN, DO, COM
5	TP, AN, DO, COM, pH
6	TP, AN, DO, COM, pH, WT
7	TP, AN, DO, COM, pH, WT, TN
8	TP, AN, DO, COM, pH, WT, TN, Tur
9	TP, AN, DO, COM, pH, WT, TN, Tur, EC

Table 6
Details of tuning parameters and settings of the optimal models for WQI predicting.

Models	Tuning parameters
LM	intercept = True
ANN	size = 5, decay = 1×10^{-4}
SOM	xdim = 3, ydim = 4, user.weights = 0.5,topo = hexagonal
KNN	kmax = 9, distance = 2, kernel = optimal
SVM	radial basis function kernel, sigma = 0.152, C = 1
RF	mtry = 5
XGB	nrounds = 150, lambda = 0, alpha = 0.1, eta = 0.3
SGB	n.trees = 150, interaction.depth = 3, shrinkage = 0.1, n.minobsinnode = 10

#All optimal models were selected by the smallest RMSE except the LM.

superiority than ANN, SVM and RF with $R^2 = 0.969/0.987$ in training and testing phase [113]. Uddin et al. [114] also indicated that KNN and XGB algorithms outperformed than other commonly used machine learning models in accuracy of WQI prediction in Cork Harbour, Ireland. In general, the decision tree-based or ensemble trees algorithms outperformed statistical and neural network models due to their ensemble learning ability from multiple weak learning machines or simple trees, which was consistent with previous studies [4,115].

A model could be considered satisfactory with R^2 higher than 0.9 in both the training and testing phases based on recommended from Grassi et al. [116]. The LM and SOM with all input groups failed to meet this criterion, while the ANN, KNN, SVM and SGB with four water quality parameters (TP, AN, DO, COM) could meet it. The RF and XGB had better performance, with R^2 values in the training and testing phases higher than 0.9 with three independent variables (TP, AN and DO) to predict the WQI. The comparisons between actual and predicted WQI from LM and machine learning models were shown in Fig. 4. The red lines were 1:1 line and if all points located in the lines indicated the predicted values were the same as the actual ones and model could exactly describe all influence factors [117]. According to the figure, it was convenient to find that RF and XGB had the best model performance where all points were alongside the red lines and the distances between scatter points and the lines were shrunk with more variables added. The points from ANN, KNN, SCM and SGB had the similar patterns with the RF and XGB, but the points distribution strips were wider. The points from LM had a reverse L-shaped, which had no visible changes with added input variables, indicating that the LM failed to provide suitable prediction for low WQI values. The performance of SOM did not significantly improve more input variables, suggesting that it was unsuited for WQI prediction in the research area.

3.3.2. Model evaluation in prediction of WQI grades

In addition to WQI values, WQI grades are also widely applied in water quality evaluation, which had five levels: “Excellent”, “Good”, “Medium”, “Low” and “Bad” water quality. In the dataset, most records were assessed as “Medium” water quality (69.0 %) while “Good” (15.5 %) and “Low” (15.2 %) levels had similar numbers of records. No record was evaluated as “Excellent” and less than 0.5 % record as “Bad” level. The prediction accuracy based on each level and the results are summarized in Table 7. The accuracies with one variable (TP) from all models ranged from 70 % to 77 %, and the accuracies generally increased with the addition of more variables into the models. The maximum accuracies from LM and SOM were about 80 %, and they were about 93 % from ANN and KNN with nine independent variables, except SOM (seven independent variables). The accuracies from SVM, RF, XGB and SGB were even higher than 95 %. A prediction accuracy of WQI levels higher than 0.85 was considered satisfactory based on the recommendation from Freeman and Moisen [118]. Therefore, the ANN, KNN, SVM, XGB and SBG performed well enough with three input variables (TP, AN and DO). The RF had even better performance than the XGB, as only two water quality parameters (TP and AN) could provide more than 85 % accuracy in predicting WQI levels.

The models whose prediction accuracies were higher than 85 % with the least numbers of input variables in each WQI grade were further analyzed, and the results are shown in Fig. 6. Due to two input variables and three variables in RF providing more than 85 % and 90 % accuracy, respectively, the model performances for both were included in the analysis. When the WQI level was “Good”, the predicted result was about 60 % “Good” level and 40 % “Medium” level. No record was erroneously predicted to be “Excellent”, “Low” or “Bad” levels. The RF_3 (RF with three parameters) had the highest prediction accuracy in the “Good” level (76 %).

Models preformed best when the actual WQI level was “Medium”, with prediction accuracies all higher than 90 % (about 92 %–96 %) and the mistaken evaluations were mainly lying in the “Good” and “Low” levels. When the WQI level was “Low”, the prediction accuracies were mainly about 85 % while the highest accuracy (90 %) also from RF_3. The prediction performance in “Low” level was not as good as the previous levels, with the highest accuracy of 47 % from RF_3, indicating the need for further research to improve model performance in predicting “Low” water quality. Overall, the RF model had the best performance in WQI grade prediction comparing to other models applied in the research, consistent with Uddin et al. [97] which reported that decision-tree models had better performance for prediction coastal water quality.

3.3.3. Selection of key water parameters

Based on the performance of prediction models, the input variables group which could explain more than 90 % variance in WQI values or have 85 % prediction accuracy in WQI grades, could be identified as the key water parameters. Accordingly, the R^2 values from RF or XGB models with TP, AN and DO were higher than 0.9. For predicting WQI grades, an accuracy of 85 % was obtained using just TP and AN with the RF model, and TP, AN, and DO with ANN, KNN, SVM, XGB, and SGB models. Wu et al. [119] used the stepwise

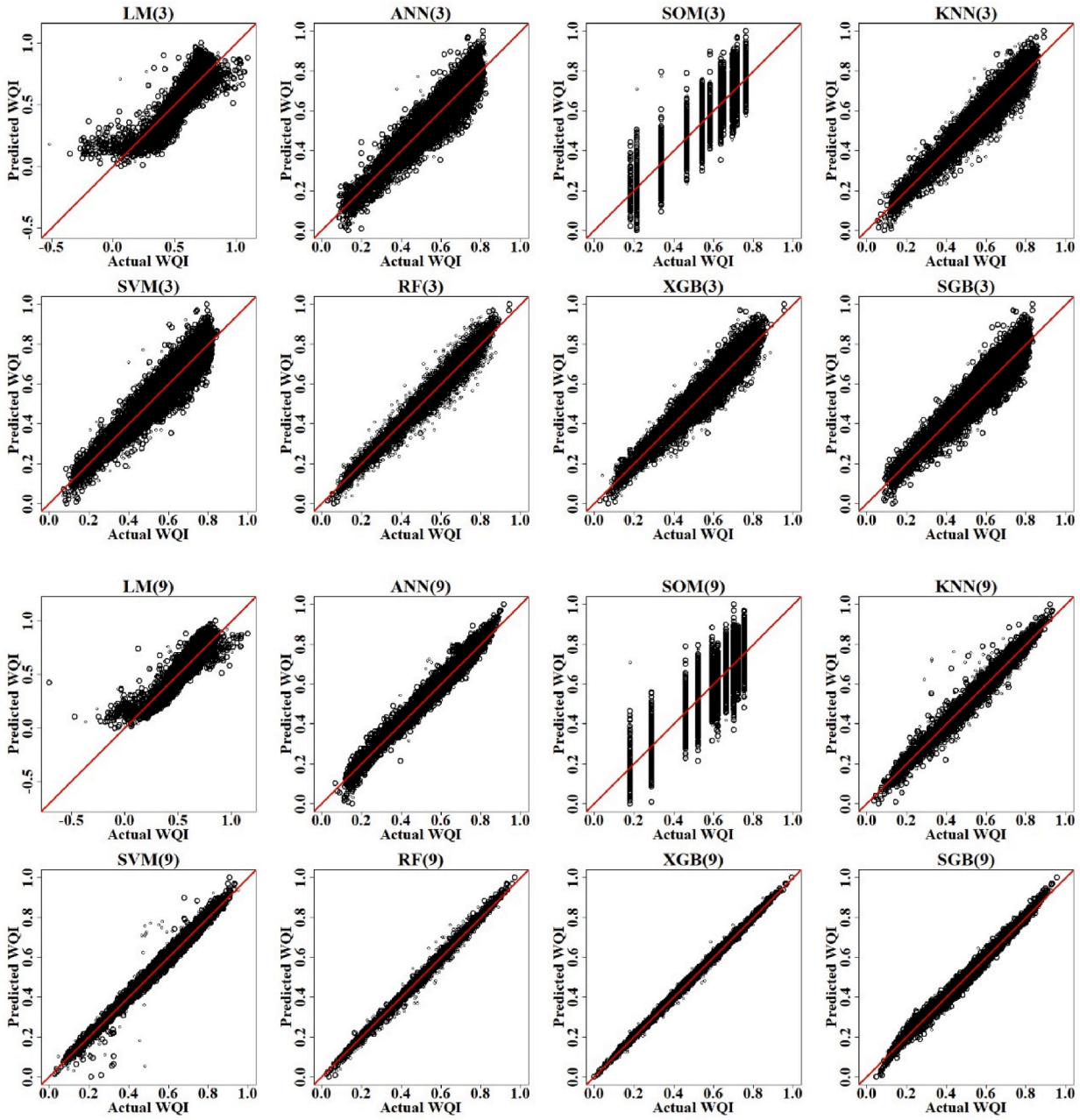

Fig. 4. The actual vs. predicted WQI which were predicted by three (key water parameters determined in the article) and nine (all water parameters) water quality parameters from different statistical and machine learning models. The comparisons between actual vs. predicted WQI from other numbers of variables were present in Fig. S1.

multiple linear regression to select six parameters from thirteen water quality indicator to calculate WQI_{min} and Yongo et al. [120] selected three key parameters from eight to calculate the WQI_{min} with the R^2 of 0.783, which indicated that the most reliable models selected in the study had superior performance in applying less number of key parameters and getting higher prediction performance. Our result was corresponding to Kisi et al. [121] which pointed out that the tree-based algorithms with no hidden layers like RF had better model performance than ANN and Chen et al. [50] proved that tree-based models had significantly better performance than other machine learning models in prediction water quality levels.

Therefore, the three key parameters TP, AN, and DO were determined for predicting WQI values using RF and XGB models, while TP and AN were the key parameters for predicting WQI grades with the RF model. Sang et al. [122] identified chlorophyll-a, DO, AN, WT, pH, and TN as key water parameters by RF models when analyzed the long-term water quality the Three Gorges Reservoir. Pan et al. [48] selected six parameters including total suspended solids, AN, COM, EC, DO, and nitrate-nitrogen to established WQI

Table 7
Accuracy of grades of water quality based on predicted WQI values from different statistical or machine learning models with nine groups of variables.

		E	G	M	L	B	Sum
Actual WQI		0	2812	12476	2741	55	18084
Model							Accuracy
LM	1	0	63	12164	927	18	72.8 %
	2	0	7	12322	872	20	73.1 %
	3	0	461	11857	1385	27	75.9 %
	4	0	863	11789	1718	35	79.7 %
	5	0	991	11771	1718	35	80.3 %
	6	0	990	11778	1722	35	80.3 %
	7	0	1007	11745	1754	35	80.4 %
	8	0	1175	11830	1767	39	81.9 %
	9	0	1259	11812	1804	37	82.5 %
ANN	1	0	654	11599	1421	0	75.6 %
	2	0	654	11644	1550	0	76.6 %
	3	0	1690	11445	2266	6	85.2 %
	4	0	1866	11571	2327	16	87.3 %
	5	0	1886	11518	2342	24	87.2 %
	6	0	1893	11565	2364	19	87.6 %
	7	0	1851	11634	2458	9	88.2 %
	8	0	2257	11851	2547	0	92.1 %
	9	0	2304	11982	2565	4	93.2 %
SOM	1	0	1277	10938	1427	0	75.4 %
	2	0	1925	10353	1495	0	76.2 %
	3	0	1009	11754	1616	0	79.5 %
	4	0	1068	10954	2514	0	80.4 %
	5	0	1210	11650	1697	0	80.5 %
	6	0	849	11538	1405	0	76.3 %
	7	0	677	11948	1970	0	80.7 %
	8	0	702	11906	1541	0	78.2 %
	9	0	925	10616	2377	0	77.0 %
KNN	1	0	462	10644	1922	0	72.0 %
	2	0	1332	11557	1680	2	80.6 %
	3	0	1936	11693	2339	12	88.4 %
	4	0	2056	11876	2429	17	90.6 %
	5	0	2080	11947	2406	15	91.0 %
	6	0	2214	11958	2447	19	92.0 %
	7	0	2266	12022	2498	18	92.9 %
	8	0	2369	12126	2492	22	94.1 %
	9	0	2358	12128	2471	23	93.9 %
SVM	1	0	707	11551	1427	0	75.7 %
	2	0	1389	11390	1468	0	78.8 %
	3	0	1825	11461	2296	11	86.2 %
	4	0	1901	11614	2378	16	88.0 %
	5	0	1901	11690	2396	20	88.5 %
	6	0	1974	11730	2396	16	89.1 %
	7	0	2035	11805	2464	16	90.2 %
	8	0	2448	12099	2570	30	94.8 %
	9	0	2461	12124	2593	33	95.2 %
RF	1	0	818	11554	1403	7	76.2 %
	2	0	1611	11865	1928	11	85.2 %
	3	0	2149	11977	2460	26	91.9 %
	4	0	2275	12084	2528	32	93.6 %
	5	0	2283	12131	2525	29	93.8 %
	6	0	2312	12186	2526	27	94.3 %
	7	0	2355	12231	2587	33	95.1 %
	8	0	2553	12350	2646	41	97.3 %
	9	0	2548	12373	2644	42	97.4 %
XGB	1	0	866	11454	1428	2	76.0 %
	2	0	1338	11572	1632	1	80.4 %
	3	0	1911	11656	2347	18	88.1 %
	4	0	2129	11876	2470	33	91.3 %
	5	0	2211	11942	2498	38	92.3 %
	6	0	2260	11983	2519	34	92.9 %
	7	0	2360	12076	2575	36	94.3 %
	8	0	2609	12317	2666	49	97.6 %
	9	0	2672	12340	2679	52	98.1 %
SGB	1	0	824	11471	1422	0	75.9 %
	2	0	1281	11482	1464	0	78.7 %

(*continued on next page*)

Table 7 (*continued*)

	E	G	M	L	B	Sum
Actual WQI	0	2812	12476	2741	55	18084
Model						Accuracy
3	0	1879	11515	2174	11	86.1 %
4	0	1943	11678	2306	20	88.2 %
5	0	1964	11733	2318	26	88.7 %
6	0	2067	11722	2323	21	89.2 %
7	0	2101	11829	2379	21	90.3 %
8	0	2538	12190	2517	28	95.5 %
9	0	2590	12266	2567	27	96.5 %

#The E, G, M, Land B mean the Excellent, Good, Medium, Low and Bad water quality, which were five levels of WQI in the research.

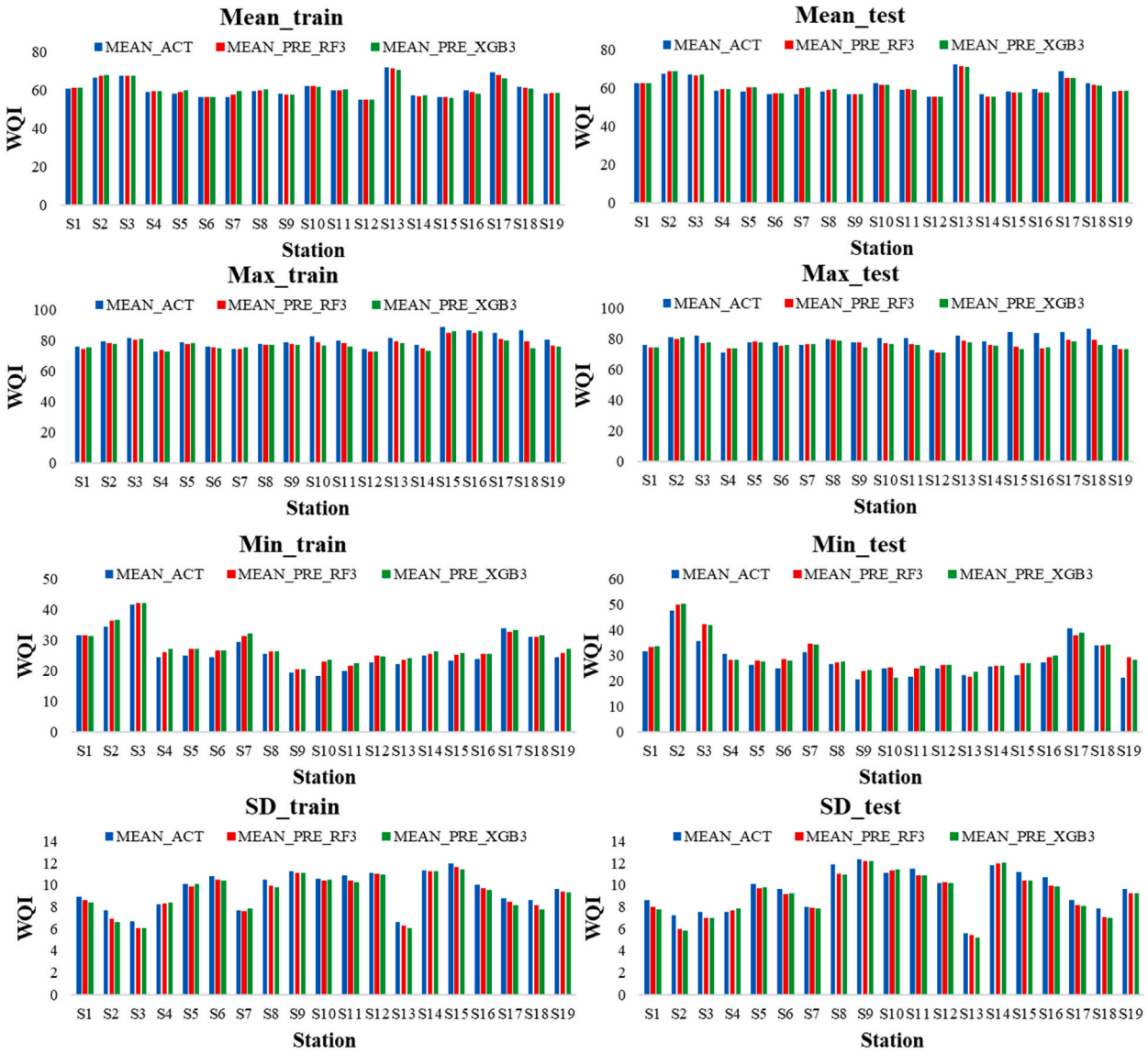

Fig. 5. The mean, maximum (MAX), Minimum (MIN) and standard deviation (SD) of actual and predicted WQI from the RF and XGB with three key water parameters.

prediction models in the Yellow River. Previous studies pointed out that AN and TP were meaningful key parameters in streams influenced by rural and agricultural activities [123,124]. The findings from Zhang et al. [125] supported the result by reporting the highest contribution to nutrients loadings from increase of build-up and agricultural land use, indicating the influence of intensive human activities exceeded the power of influence of natural processes on rivers.

3.3.4. Model uncertainty of RF and XGB

Based on the results from predicting the WQI, the RF and XGB models were selected as the most reliable, with three key water parameters identified. To assess the reliability of these models, an analysis of their uncertainty was conducted at each monitoring station (Fig. 5).

The mean of the actual and predicted WQI values were nearly identical across stations, except at S17, where the predicted mean was slightly lower than the actual mean during the testing phase. The maximum and minimum predicted values did not significantly differ from the actual values, with the exception of stations S15–S17, where the predicted maximum values were lower than the actual maximums. The SD of the predicted WQI were similar or slight lower than the actual SD, indicating that the range of predicted WQI values was slightly narrower than the actual range. This narrower range is consistent with the finding that the model prediction accuracy was around 50 % for the "Low" water quality grade, suggesting directions for further model improvement. Overall, however, the predicted WQI values from the RF and XGB models using the key parameters were reliable when compared to the actual WQI values.

3.4. Spatio-temporal variations of key water parameters

The spatio-temporal patterns of the three key water parameters were present in Fig. 7. Generally, TP concentration increased from inland to coastal stations (Fig. 7 (a)). Nantong exhibited more severe TP pollution compared to Yancheng, with the most polluted areas located along the regional borders of the two cities (stations S9, S10, S11 in Yancheng and S12, S14, S15 in Nantong). Although some stations showed relatively low TP levels in 2022, the overall trend was an increase over the study period. According to water resources bulletins, the lower annual rainfall in 2022 (407 mm and 180 mm less than 2021 in the two cities, respectively) suggests that TP pollution likely originated from non-point sources, such as fertilizer and pesticide runoff from farmlands, and livestock and poultry manure [126–129].

AN concentration exhibited a similar spatial pattern to TP, increasing from inland to coastal stations (Fig. 7 (b)). This finding was corresponding to results from Chabuk et al. [130] which pointed out area downstream of the Tigris River was more polluted. Among

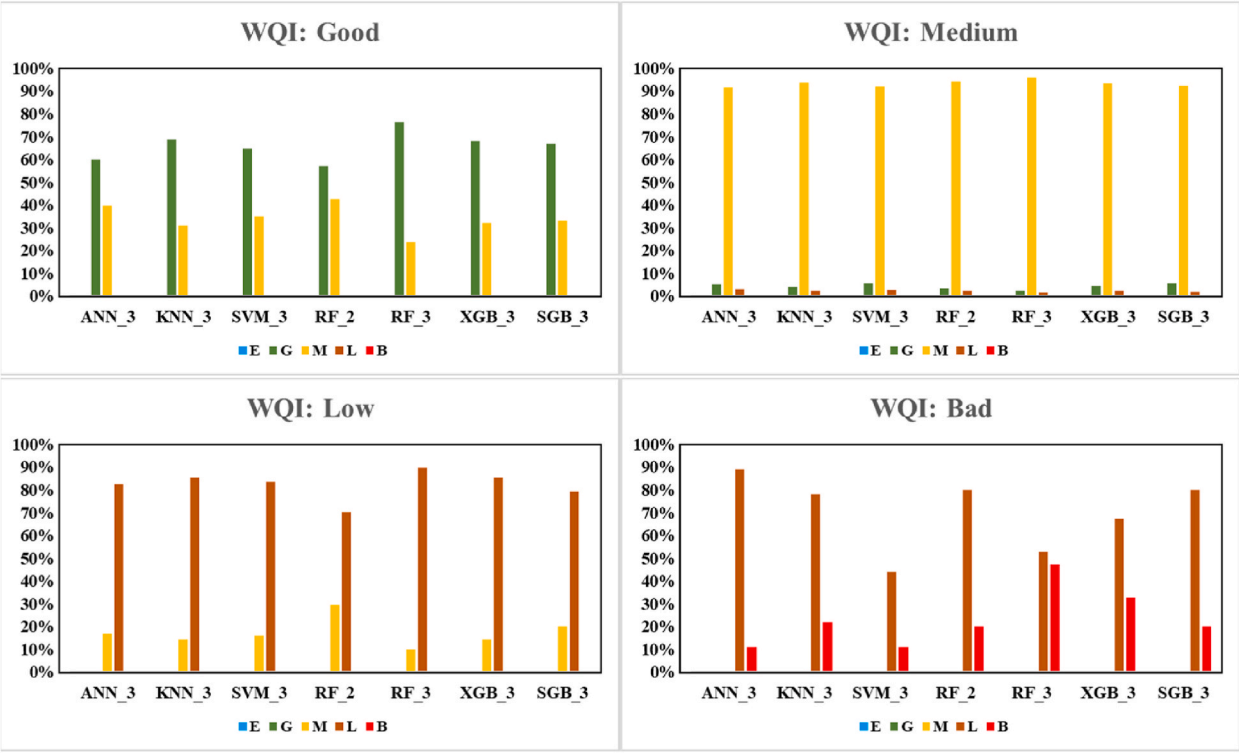

Fig. 6. The prediction accuracy of models which obtained 85 % accuracy in prediction of WQI classification with the least water quality parameters. The ANN, KNN, SVM, XGB and SGB with three water quality parameters obtained 85 % accuracy. The prediction accuracies from RF with two and three water quality parameters were higher than 85 % and 90 % respectively. Due to no excellent WQI existed in the dataset, four rating of WQI (good, medium, low and bad) were shown in the figures.

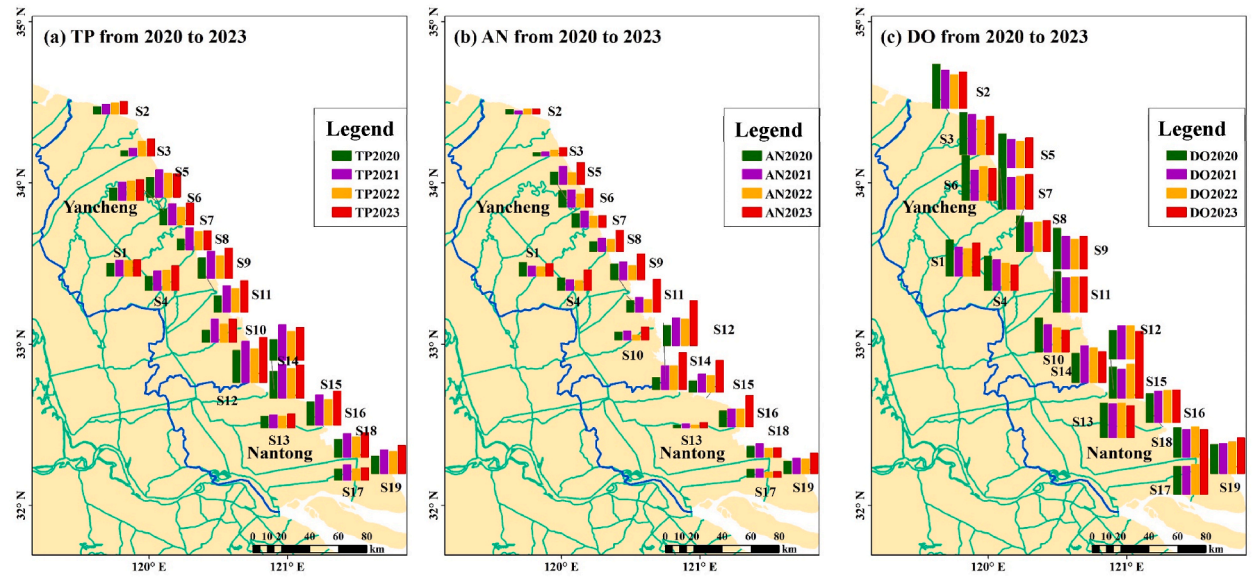

Fig. 7. The variations of key water quality parameters (TP, AN, DO) and the WQI calculated by all water quality parameters from 2020 to 2023 at nineteen monitor stations.

the coastal stations, the most polluted areas were in the central region and northern part of Yancheng (S2 and S3) and the southern part of Nantong (S17, S18, and S19). While AN concentration was relatively stable from 2020 to 2022, nearly all stations witnessed a 60 % increase in 2023, highlighting the need for river management efforts to control AN pollution.

DO concentrations in the southern stations were slightly higher than in the northern stations, suggesting better self-purification capacity of the water bodies in Nantong compared to Yancheng (Fig. 7 (c)). Contrary to the spatial variations of TP and AN, DO concentrations were lower in upstream stations (S1 and S4) than in coastal stations. DO levels decreased from 2020 to 2022 and increased in 2023 at most stations, except for S4 and S10 in Yancheng, which exhibited a continued decline, while DO remained stable or slightly increased over the study period in Nantong.

3.5. Implications and limitations

The findings of this research have significant practical and policy implications for river protection in coastal cities of Jiangsu Province, and could also be applicable to other regions with similar natural processes and anthropogenic disturbances. Accordingly, the results contribute to acceptable prediction of river water environment levels in Yancheng and Nantong cities using just two or three water quality parameters and machine learning models. This could help local governments take quick measures to control pollutants or jointly regulate water quality and quantity during sudden pollution incidents [131,132]. The scientific implications lie in promoting the advantages of machine learning models for environmental evaluation, especially when data availability is limited due to discontinuous or insufficient monitoring caused by time or funding constraints. This research aims to further propel the development and application of artificial intelligence in river protection.

This research provides clear insights that nutrients are main factors affecting water quality in these coastal cities adjacent to the Yellow Sea. This is supported by Rao et al. [72] who reports of high nutrients concentrations and loads in rivers near rural land in Yancheng, which still exceed national standard, emphasizing the need for continuous monitoring to combat water pollution and ensure water safety. By determining key water parameters using machine learning models, not only can the cost of water quality monitoring be effectively reduced, but also valuable guidance can be provided for agricultural practices and industrial production to ensure sustainable development of the water environment.

In this study, RF and XGB models demonstrated excellent performance in predicting WQI values and grades. However, the prediction accuracy for the “Bad” level (<50 %) was not satisfactory compared to other levels (“Good”, “Medium” and “Low” water quality). This could be due to the small quantity of data (<0.3 %) or relatively low WQI values in this category. To further improve model performance for the "Low" water quality level, which is crucial for identifying and providing early warning of water pollution incidents, more water quality data should be collected, and model improvements should be made.

4. Conclusions

In this study, the WQI was efficiently assessed and predicted using key water parameters and reliable models for two coastal cities, Yancheng and Nantong, adjacent to the Yellow Sea in Jiangsu province, China. The key findings are.

(1) Water quality improved from 2020 to 2022 but deteriorated in 2023. Inland stations exhibited better water quality than coastal stations, particularly in terms of turbidity and nutrient levels. The water environment in Nantong was relatively better than Yancheng, with mean WQI values of 55.3–72.0 and 56.4–67.3, respectively. TN was the most serious pollution, failing to meet even the Grade V criteria.
(2) "Good" and "Medium" water quality classifications accounted for 80 % of the stations. The highest "Good" percentage was 71 % (S13, Nantong), while the lowest was about 1 % (S4, S7 in Yancheng and S12 in Nantong). No station achieved "Excellent" level, and "Bad" level was recorded at less than 2 %. TP showed the strongest negative correlation (−0.72) with WQI, followed by AN, DO, COM and pH, all of which had correlations above 0.5.
(3) Performance of all model in predicting WQI values improved with the addition of input variables, except for SOM. The NSE and R^2 values were even higher than 0.99 with nine input variables from machine learning models such as SVM, RF, XGB and SGB. The satisfactory prediction models with the minimum key parameters were RF and XGB with TP, AN and DO, whose NSE and R^2 values were higher than 0.9.
(4) For predicting WQI grades, SVM, RF, XGB, and SGB achieved 95 % accuracy with nine inputs. RF performed best, with over 85 % accuracy using just TP and AN. Accuracies exceeded 90 % for "Medium" and "Low" grades and around 70 % for "Good", but below 50 % for "Bad" grade with TP, AN, and DO.

The model approaches to determine the most reliable machine learning models and key water parameters could also be appliable in other rivers. This could contribute to quicker WQI predictions by saving costs and reducing time spent on sample collection and laboratory analysis, thereby facilitating improvements in water resources management and pollution control in rivers. Model performance in predicting of bad water quality conditions still had a margin to be satisfactory in the article, thus, further research should focus on modifying model structures to address the limitation. Although records of poor water conditions, such as sudden pollution incidents, occupy a small portion of the data, accurate predictions are crucial for the protection of the water environment. Moreover, seasonal changes in water quality may impact WQI prediction; therefore, developing seasonal WQI prediction models to explore the influence of seasonal variations will be a focus of further research.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Jing Xu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Yuming Mo:** Writing – review & editing, Investigation. **Senlin Zhu:** Writing – review & editing, Methodology. **Jinran Wu:** Writing – review & editing, Methodology. **Guangqiu Jin:** Writing – review & editing. **You-Gan Wang:** Validation, Methodology, Investigation. **Qingfeng Ji:** Software, Methodology. **Ling Li:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing 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.

Acknowledgments

This research was supported by the Belt and Road Special Foundation of The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing, China (2021491811, 2022491411).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e33695.

References

[1] D.T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, N. Kazakis, Improving prediction of water quality indices using novel hybrid machine-learning algorithms, Sci. Total Environ. 721 (2020/06/15/2020) 137612.
[2] Q. Wang, Z. Li, Y. Xu, R. Li, M. Zhang, Adaptive-weight water quality assessment and human health risk analysis for river water in Hong Kong, Environ. Sci. Pollut. Control Ser. 29 (2022/10/01 2022) 75936–75954.
[3] C.J. Vörösmarty, P.B. McIntyre, M.O. Gessner, D. Dudgeon, A. Prusevich, P. Green, et al., Global threats to human water security and river biodiversity, Nature 467 (2010/09/01 2010) 555–561.
[4] H. Lu, X. Ma, Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere 249 (2020/06/01/2020) 126169.

[5] M. Akbari, F. Samadzadegan, R. Weibel, A generic regional spatio-temporal co-occurrence pattern mining model: a case study for air pollution, J. Geogr. Syst. 17 (Jul 2015) 249–274.
[6] Y. Cui, J. Dong, H. Wang, M. Shang, H. Xie, Y. Du, et al., Spatiotemporal response of water quality in fragmented mangroves to anthropogenic activities and recommendations for restoration, Environ. Res. 237 (2023) 117075.
[7] N. Bansal, Industrial development and challenges of water pollution in coastal areas: the case of Surat, India, IOP Conf. Ser. Earth Environ. Sci. 120 (2018/03/01 2018) 012001.
[8] V.M. Gruzinov, N.N. Dyakov, I.V. Mezenceva, Y.A. Malchenko, N.V. Zhohova, A.N. Korshenko, Sources of coastal water pollution near Sevastopol, Oceanology 59 (2019/07/01 2019) 523–532.
[9] D.I. Taylor, C.A. Oviatt, A.E. Giblin, J. Tucker, R.J. Diaz, K. Keay, Wastewater input reductions reverse historic hypereutrophication of Boston Harbor, USA, Ambio 49 (2020/01/01 2020) 187–196.
[10] G. Qu, An evaluation method for water pollution Treatment efficiency in coastal cities based on regional management, J. Coast Res. (2020) 100–103.
[11] M.-L. Wu, Y.-S. Wang, C.-C. Sun, H. Wang, J.-D. Dong, J.-P. Yin, et al., Identification of coastal water quality by statistical analysis methods in Daya Bay, South China Sea, Mar. Pollut. Bull. 60 (2010/06/01/2010) 852–860.
[12] S. Liu, S. Lou, C. Kuang, W. Huang, W. Chen, J. Zhang, et al., Water quality assessment by pollution-index method in the coastal waters of Hebei Province in western Bohai Sea, China, Mar. Pollut. Bull. 62 (2011/10/01/2011) 2220–2229.
[13] M. Zhang, X. Sun, J. Xu, Heavy metal pollution in the East China Sea: a review, Mar. Pollut. Bull. 159 (2020/10/01/2020) 111473.
[14] W. Shang, M. Yang, Z. Han, X. Chen, Distribution, contamination assessment, and sources of heavy metals in surface sediments from the south of the North Yellow Sea, China, Mar. Pollut. Bull. 196 (Nov 2023).
[15] Z. Dong, D. Liu, Y. Wang, B. Di, Temporal and spatial variations of coastal water quality in Sishili Bay, northern Yellow Sea of China, Aquat. Ecosys. Health Manag. 22 (2019) 30–39.
[16] X. Sun, Z. Dong, W. Zhang, X. Sun, C. Hou, Y. Liu, et al., Seasonal and spatial variations in nutrients under the influence of natural and anthropogenic factors in coastal waters of the northern Yellow Sea, China, Mar. Pollut. Bull. 175 (2022/02/01/2022) 113171.
[17] J. Wang, M. Wang, S. Ru, X. Liu, High levels of microplastic pollution in the sediments and benthic organisms of the South Yellow Sea, China, Sci. Total Environ. 651 (2019/02/15/2019) 1661–1669.
[18] Q.F. Han, S. Zhao, X.R. Zhang, X.L. Wang, C. Song, S.G. Wang, Distribution, combined pollution and risk assessment of antibiotics in typical marine aquaculture farms surrounding the Yellow Sea, North China, Environ. Int. 138 (2020/05/01/2020) 105551.
[19] L. Zhu, H. Bai, B. Chen, X. Sun, K. Qu, B. Xia, Microplastic pollution in North Yellow Sea, China: observations on occurrence, distribution and identification, Sci. Total Environ. 636 (2018/09/15/2018) 20–29.
[20] F. Xiong, Y. Chen, S. Zhang, Y. Xu, Y. Lu, X. Qu, et al., Land use, hydrology, and climate influence water quality of China?s largest river, J. Environ. Manag. 318 (Sep 15 2022).
[21] K. Tian, Q. Wu, P. Liu, W. Hu, B. Huang, B. Shi, et al., Ecological risk assessment of heavy metals in sediments and water from the coastal areas of the Bohai Sea and the Yellow Sea, Environ. Int. 136 (2020/03/01/2020) 105512.
[22] K.B. Githaiga, S.M. Njuguna, R.W. Gituru, X. Yan, Water quality assessment, multivariate analysis and human health risks of heavy metals in eight major lakes in Kenya, J. Environ. Manag. 297 (Nov 1 2021).
[23] S. Giri, Z. Qiu, Understanding the relationship of land uses and water quality in Twenty First Century: a review, J. Environ. Manag. 173 (May 15 2016) 41–48.
[24] S. Xu, Y. Cui, C. Yang, S. Wei, W. Dong, L. Huang, et al., The fuzzy comprehensive evaluation (FCE) and the principal component analysis (PCA) model simulation and its applications in water quality assessment of Nansi Lake Basin, China, Environmental Engineering Research 26 (2021).
[25] H. Gharibi, A.H. Mahvi, R. Nabizadeh, H. Arabalibeik, M. Yunesian, M.H. Sowlat, A novel approach in water quality assessment based on fuzzy logic, J. Environ. Manag. 112 (Dec 15 2012) 87–95.
[26] G. Hu, H.R. Mian, Z. Abedin, J. Li, K. Hewage, R. Sadiq, Integrated probabilistic-fuzzy synthetic evaluation of drinking water quality in rural and remote communities, J. Environ. Manag. 301 (Jan 1 2022).
[27] S. Elsayed, H. Ibrahim, H. Hussein, O. Elsherbiny, A.H. Elmetwalli, F.S. Moghanm, et al., Assessment of water quality in lake Qaroun using ground-based remote sensing data and artificial neural networks, Water 13 (2021) 3094.
[28] M. Gad, A.H. Saleh, H. Hussein, M. Farouk, S. Elsayed, Appraisal of surface water quality of nile river using water quality indices, Spectral signature and multivariate modeling, Water 14 (2022) 1131.
[29] T. Yan, S.-L. Shen, A. Zhou, Indices and models of surface water quality assessment: review and perspectives, Environ. Pollut. 308 (Sep 1 2022).
[30] C. Benaissa, B. Bouhmadi, A. Rossi, An assessment of the physicochemical, bacteriological quality of groundwater and the water quality index (WQI) used GIS in Ghis Nekor, Northern Morocco, Scientific African 20 (2023/07/01/2023) e01623.
[31] C. Benaissa, B. Bouhmadi, A. Rossi, Y. El Hammoudani, F. Dimane, Assessment of water quality using water quality index – case study of bakoya aquifer, Al hoceima, northern Morocco, Ecol. Eng. Environ. Technol. 23 (2022 2022) 31–44.
[32] A.W. Brown, The role of government in the encouragement of research in industry, Aust. J. Publ. Adm. 29 (1970) 339–355.
[33] M.G. Uddin, S. Nash, A.I. Olbert, A review of water quality index models and their use for assessing surface water quality, Ecol. Indicat. 122 (2021/03/01/2021) 107218.
[34] K. Cash, R. Wright, Canadian Water Quality Guidelines for the Protection of Aquatic Life, CCME, Ottawa, ON, Canada, 2001.
[35] A. Lumb, T. Sharma, J.-F. Bibeault, A review of genesis and evolution of water quality index (WQI) and some future directions, Water Quality 3 (2011) 11–24.
[36] M.G. Uddin, S. Nash, A. Rahman, A.I. Olbert, A sophisticated model for rating water quality, Sci. Total Environ. 868 (2023).
[37] Z. Wu, X. Wang, Y. Chen, Y. Cai, J. Deng, Assessing river water quality using water quality index in Lake Taihu Basin, China, Sci. Total Environ. 612 (2018) 914–922.
[38] W. Sun, C. Xia, M. Xu, J. Guo, G. Sun, Application of modified water quality indices as indicators to assess the spatial and temporal trends of water quality in the Dongjiang River, Ecol. Indicat. 66 (2016) 306–312.
[39] Z. Wu, X. Lai, K. Li, Water quality assessment of rivers in Lake Chaohu Basin (China) using water quality index, Ecol. Indicat. 121 (2021) 107021.
[40] Y. Tian, Y. Jiang, Q. Liu, M. Dong, D. Xu, Y. Liu, et al., Using a water quality index to assess the water quality of the upper and middle streams of the Luanhe River, northern China, Sci. Total Environ. 667 (2019) 142–151.
[41] P.R. Kannel, S. Lee, Y.-S. Lee, S.R. Kanel, S.P. Khan, Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment, Environ. Monit. Assess. 132 (2007/09/01 2007) 93–110.
[42] S.F. Pesce, D.A. Wunderlin, Use of water quality indices to verify the impact of Córdoba City (Argentina) on Suquia River, Water Res. 34 (2000/08/01/2000) 2915–2926.
[43] P.-L. Georgescu, S. Moldovanu, C. Iticescu, M. Calmuc, V. Calmuc, C. Topa, et al., Assessing and forecasting water quality in the Danube River by using neural network approaches, Sci. Total Environ. 879 (2023).
[44] A.A. Bordalo, R. Teixeira, W.J. Wiebe, A water quality index applied to an international shared River Basin: the case of the douro river, Environ. Manag. 38 (2006/12/01 2006) 910–920.
[45] X. Nong, D. Shao, H. Zhong, J. Liang, Evaluation of water quality in the South-to-North Water Diversion Project of China using the water quality index (WQI) method, Water Res. 178 (2020) 115781.
[46] J. Wu, S.-P. Cheng, L.-Y. He, Y.-C. Wang, Y. Yue, H. Zeng, et al., Assessing water quality in the Pearl River for the last decade based on clustering: characteristic, evolution and policy implications, Water Res. 244 (2023/10/01/2023) 120492.
[47] J. Chen, T. Yang, Y. Wang, H. Jiang, C. He, Effects of ecological restoration on water quality and benthic macroinvertebrates in rural rivers of cold regions: a case study of the Huaide River, Northeast China, Ecol. Indicat. 142 (2022/09/01/2022) 109169.
[48] B. Pan, X. Han, Y. Chen, L. Wang, X. Zheng, Determination of key parameters in water quality monitoring of the most sediment-laden Yellow River based on water quality index, Process Saf. Environ. Protect. 164 (2022/08/01/2022) 249–259.

[49] J. Qi, L. Yang, E. Liu, A holistic framework of water quality evaluation using water quality index (WQI) in the Yihe River (China), Environ. Sci. Pollut. Control Ser. 29 (2022) 80937–80951.
[50] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, et al., Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res. 171 (2020/03/15) (2020) 115454.
[51] Q. Wang, Z. Li, J. Cai, M. Zhang, Z. Liu, Y. Xu, et al., Spatially adaptive machine learning models for predicting water quality in Hong Kong, J. Hydrol. 622 (2023/07/01/2023) 129649.
[52] C. Zhang, Y. Lu, Study on artificial intelligence: the state of the art and future prospects, Journal of Industrial Information Integration 23 (2021/09/01/2021) 100224.
[53] N. Ebadati, M. Hooshmandzadeh, Water quality assessment of river using RBF and MLP methods of artificial network analysis (case study: karoon River Southwest of Iran), Environ. Earth Sci. 78 (2019/08/29 2019) 551.
[54] M. Kulisz, J. Kujawska, Application of artificial neural network (ANN) for water quality index (WQI) prediction for the river Warta, Poland, J. Phys. Conf. 2130 (2021/12/01 2021) 012028.
[55] M. Najafzadeh, A. Ghaemi, Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods, Environ. Monit. Assess. 191 (2019) 1–21.
[56] S. Khullar, N. Singh, Machine learning techniques in river water quality modelling: a research travelogue, Water Supply 21 (2020) 1–13.
[57] A.H. Haghiabi, A.H. Nasrolahi, A. Parsaie, Water quality prediction using machine learning methods, Water Quality Research Journal 53 (2018) 3–13.
[58] T. M. Tung Tiyasha, Z.M. Yaseen, A survey on river water quality modelling using artificial intelligence models: 2000–2020, J. Hydrol. 585 (2020/06/01/2020) 124670.
[59] J. Gao, G. Deng, H. Jiang, Y. Wen, S. Zhu, C. He, et al., Water quality pollution assessment and source apportionment of lake wetlands: a case study of Xianghai Lake in the Northeast China Plain, J. Environ. Manag. 344 (2023) 118398.
[60] Q.-w. Yu, F.-p. Wu, Z.-f. Zhang, Z.-c. Wan, J.-y. Shen, L.-n. Zhang, Technical inefficiency, abatement cost and substitutability of industrial water pollutants in Jiangsu Province, China, J. Clean. Prod. 280 (2021) 124260.
[61] Y. Kong, W. He, J. Shen, L. Yuan, X. Gao, T.S. Ramsey, et al., Adaptability analysis of water pollution and advanced industrial structure in Jiangsu Province, China, Ecol. Model. 481 (2023) 110365.
[62] B. Zuo, C. Liu, R. Chen, H. Kan, J. Sun, J. Zhao, et al., Associations between short-term exposure to fine particulate matter and acute exacerbation of asthma in Yancheng, China, Chemosphere 237 (2019/12/01/2019) 124497.
[63] H. Yao, Characterizing landuse changes in 1990–2010 in the coastal zone of Nantong, Jiangsu province, China, Ocean Coast Manag. 71 (2013/01/01/2013) 108–115.
[64] J. Wang, Y. Mao, J. Xu, c. Jiang, Groundwater quality and cause analysis in Yancheng area, Jiangsu Province, J. Anhui Agric. Univ. 50 (2023) 116–125.
[65] Z. Chen, H. Zhang, M. Xu, Y. Liu, J. Fang, X. Yu, et al., A study on the ecological zoning of the Nantong coastal zone based on the Marxan model, Ocean Coast Manag. 229 (2022/10/01/2022) 106328.
[66] X. Tang, M. Shen, Y. Zhang, D. Zhu, H. Wang, Y. Zhao, et al., The changes in antibiotic resistance genes during 86 years of the soil ripening process without anthropogenic activities, Chemosphere 266 (2021/03/01/2021) 128985.
[67] Y. Wang, P. Qian, D. Li, H. Chen, X. Zhou, Assessing risk to human health for heavy metal contamination from public point utility through ground dust: a case study in Nantong, China, Environ. Sci. Pollut. Control Ser. 28 (2021/12/01 2021) 67234–67247.
[68] F. Zhou, H. Lv, C. Liu, Change of river system in the Lixiahe region during urbanization, South-to-North Water Transfers and Water Science & Technology 16 (Jan 2018) 144–150.
[69] X. Deng, Distribution, Source and Response of Human Activities to Polycyclic Aromatic Hydrocarbons in Sediments of Rivers in Yancheng City," Master, 2022.
[70] G. Huang, Analysis on the distribution and evolution law of salinity in shallow groundwater (10m to shallow) in Yancheng City, Jiangsu Water Resources (2019) 8–10.
[71] Y. Chen, G. Zhu, J. Ni, Hydrochemical characteristics and water quality evaluation of shallow groundwater in Nantong urban area, Ground Water 43 (2021) 29–31+41.
[72] P. Rao, S. Wang, A. Wang, D. Yang, L. Tang, Spatiotemporal characteristics of nonpoint source nutrient loads and their impact on river water quality in Yancheng city, China, simulated by an improved export coefficient model coupled with grid-based runoff calculations, Ecol. Indicat. 142 (2022/09/01/2022) 109188.
[73] L. Ye, X. Wu, D. Yan, B. Yang, T. Zhang, D. Huang, Dissolved organic carbon content is lower in warm seasons and neutral sugar composition indicates its degradation in a large subtropical river (Nantong Section), China, Environ. Earth Sci. 78 (Mar 2019).
[74] Z. Shang, T. Deng, J. He, X. Duan, A novel model for hourly PM2.5 concentration prediction based on CART and EELM, Sci. Total Environ. 651 (2019/02/15/2019) 3043–3052.
[75] Y. Wang, R. Yu, G. Zhu, Evaluation of physicochemical characteristics in drinking water sources emphasized on fluoride: a case study of Yancheng, China, Int. J. Environ. Res. Publ. Health 16 (Mar 2 2019).
[76] R. Makubura, D.P.P. Meddage, H.M. Azamathulla, M. Pandey, U. Rathnayake, A simplified mathematical formulation for water quality index (WQI): a case study in the kelani River Basin, Sri Lanka, Fluid 7 (2022) 147.
[77] K.D. Siriwardhana, D.I. Jayaneththi, R.D. Herath, R.K. Makumbura, H. Jayasinghe, M.B. Gunathilake, et al., A simplified equation for calculating the water quality index (WQI), kalu river, Sri Lanka, Sustainability 15 (2023) 12012.
[78] J.M. Geetha, Secure water quality prediction system using machine learning and blockchain technologies, J. Environ. Manag. 350 (2024/01/15/2024) 119357.
[79] S. Tong, W. Li, J. Chen, R. Xia, J. Lin, Y. Chen, et al., A novel framework to improve the consistency of water quality attribution from natural and anthropogenic factors, J. Environ. Manag. 342 (Sep 15 2023).
[80] J. Anmala, T. Venkateshwarlu, Statistical assessment and neural network modeling of stream water quality observations of Green River watershed, KY, USA, Water Supply 19 (2019) 1831–1840.
[81] T. Mitrović, D. Antanasijević, S. Lazović, A. Perić-Grujić, M. Ristić, Virtual water quality monitoring at inactive monitoring sites using Monte Carlo optimized artificial neural networks: a case study of Danube River (Serbia), Sci. Total Environ. 654 (2019/03/01/2019) 1000–1009.
[82] J. Shen, Q. Qin, Y. Wang, M. Sisson, A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading, Ecol. Model. 398 (2019/04/24/2019) 44–54.
[83] Z.M. Yaseen, S.O. Sulaiman, R.C. Deo, K.-W. Chau, An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction, J. Hydrol. 569 (2019/02/01/2019) 387–408.
[84] M.-L. Zhang, Z.-H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recogn. 40 (2007/07/01/2007) 2038–2048.
[85] Q. Gu, H. Hu, L. Ma, L. Sheng, S. Yang, X. Zhang, et al., Characterizing the spatial variations of the relationship between land use and surface water quality using self-organizing map approach, Ecol. Indicat. 102 (2019/07/01/2019) 633–643.
[86] T. Kohonen, Exploration of large document collections by self-organizing maps, in: Proceedings of the Sixth Scandinavian Conference on Artificial Intelligence, 1998.
[87] B. Sakaa, A. Elbeltagi, S. Boudibi, H. Chaffaï, A.R.M.T. Islam, L.C. Kulimushi, et al., Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin, Environ. Sci. Pollut. Control Ser. 29 (2022/07/01 2022) 48491–48508.
[88] A.O. Alnahit, A.K. Mishra, A.A. Khan, Stream water quality prediction using boosted regression tree and random forest models, Stoch. Environ. Res. Risk Assess. 36 (2022/09/01 2022) 2661–2680.
[89] J. Velthoen, C. Dombry, J.-J. Cai, S. Engelke, Gradient boosting for extreme quantile regression, Extremes 26 (2023/12/01 2023) 639–667.
[90] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378.

[91] U. Rasool, X. Yin, Z. Xu, M.A. Rasool, V. Senapathi, M. Hussain, et al., Mapping of groundwater productivity potential with machine learning algorithms: a case study in the provincial capital of Baluchistan, Pakistan, Chemosphere 303 (2022/09/01/2022) 135265.
[92] M.G. Uddin, S. Nash, M.T. Mahammad Diganta, A. Rahman, A.I. Olbert, Robust machine learning algorithms for predicting coastal water quality index, J. Environ. Manag. 321 (2022/11/01/2022) 115923.
[93] M.G. Uddin, A. Rahman, S. Nash, M.T.M. Diganta, A.M. Sajib, M. Moniruzzaman, et al., Marine waters assessment using improved water quality model incorporating machine learning approaches, J. Environ. Manag. 344 (2023/10/15/2023) 118368.
[94] K.S. Parmar, R. Bhardwaj, Water quality management using statistical analysis and time-series prediction model, Appl. Water Sci. 4 (2014/12/01 2014) 425–434.
[95] M.G. Uddin, S. Nash, A. Rahman, A.I. Olbert, A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment, Water Res. 219 (2022/07/01/2022) 118532.
[96] H. McCuen Richard, Z. Knight, A.G. Cutter, Evaluation of the nash–sutcliffe efficiency index, J. Hydrol. Eng. 11 (2006/11/01 2006) 597–602.
[97] M.G. Uddin, S. Nash, A. Rahman, A.I. Olbert, Assessing optimization techniques for improving water quality model, J. Clean. Prod. 385 (2023/01/20/2023) 135671.
[98] M.G. Uddin, S. Nash, A. Rahman, A.I. Olbert, A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches, Water Res. 229 (2023/02/01/2023) 119422.
[99] M.G. Uddin, M.T.M. Diganta, A.M. Sajib, A. Rahman, S. Nash, T. Dabrowski, et al., Assessing the impact of COVID-19 lockdown on surface water quality in Ireland using advanced Irish water quality index (IEWQI) model, Environ. Pollut. 336 (2023/11/01/2023) 122456.
[100] A.M. Moreno-Rodenas, F. Tscheikner-Gratl, J.G. Langeveld, F.H.L.R. Clemens, Uncertainty analysis in a large-scale water quality integrated catchment modelling study, Water Res. 158 (2019/07/01/2019) 46–60.
[101] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.
[102] O.D. Ansa-Asare, I.L. Marr, M.S. Cresser, Evaluation of modelled and measured patterns of dissolved oxygen in a freshwater lake as an indicator of the presence of biodegradable organic pollution, Water Res. 34 (2000/03/01/2000) 1079–1088.
[103] D.M. Lawler, G.E. Petts, I.D.L. Foster, S. Harper, Turbidity dynamics during spring storm events in an urban headwater river system: the Upper Tame, West Midlands, UK, Sci. Total Environ. 360 (2006/05/01/2006) 109–126.
[104] R.J. Davies-Colley, D.G. Smith, Turbidity suspended sediment, and water clarity: a REVIEW1, JAWRA Journal of the American Water Resources Association 37 (2001/10/01 2001) 1085–1101.
[105] M. Xiang, Analysis of sediment transport process and sedimentation characteristics under Xinyang Port gate in Yancheng City, in: 2013 CHES Annual Conference, 2013, pp. 1079–1088.
[106] J. Wang, A.H.W. Beusen, X. Liu, R. Van Dingenen, F. Dentener, Q. Yao, et al., Spatially explicit inventory of sources of nitrogen inputs to the Yellow Sea, east China sea, and south China sea for the period 1970–2010, Earth's Future 8 (2020).
[107] R. Ahmad, D. Kim, An extended self-organizing map based on 2-opt algorithm for solving symmetrical traveling salesperson problem, Neural Comput. Appl. 26 (2015/05/01 2015) 987–994.
[108] M.J. Crespo-Ramos, I. Machón-González, H. López-García, J.L. Calvo-Rolle, Detection of locally relevant variables using SOM–NG algorithm, Eng. Appl. Artif. Intell. 26 (2013/09/01/2013) 1992–2000.
[109] B.Q. Lap, T.-T.-H. Phan, H.D. Nguyen, L.X. Quang, P.T. Hang, N.Q. Phi, et al., Predicting Water Quality Index (WQI) by feature selection and machine learning: a case study of an Kim Hai irrigation system, Ecol. Inf. 74 (2023/05/01/2023) 101991.
[110] T. M. Tung Tiyasha, Z.M. Yaseen, Deep learning for prediction of water quality index classification: tropical catchment environmental assessment, Nat. Resour. Res. 30 (2021/12/01 2021) 4235–4254.
[111] S. Kouadri, A. Elbeltagi, A.R.M.T. Islam, S. Kateb, Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast), Appl. Water Sci. 11 (2021/11/06 2021) 190.
[112] D.N. Khoi, N.T. Quan, D.Q. Linh, P.T.T. Nhi, N.T.D. Thuy, Using machine learning models for predicting the water quality index in the La Buong river, Vietnam, Water 14 (2022) 1552.
[113] U. Mohseni, C.B. Pande, S. Chandra Pal, F. Alshehri, Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model, Chemosphere 352 (2024/03/01/2024) 141393.
[114] M.G. Uddin, S. Nash, A. Rahman, A.I. Olbert, Performance analysis of the water quality index model for predicting water state using machine learning techniques, Process Saf. Environ. Protect. 169 (2023/01/01/2023) 808–828.
[115] S. Wang, H. Peng, S. Liang, Prediction of estuarine water quality using interpretable machine learning approach, J. Hydrol. 605 (2022/02/01/2022) 127320.
[116] L. Grassi, E. Schileo, F. Taddei, L. Zani, M. Juszczyk, L. Cristofolini, et al., Accuracy of finite element predictions in sideways load configurations for the proximal human femur, J. Biomech. 45 (2012/01/10/2012) 394–399.
[117] R. Pany, A. Rath, P.C. Swain, Water quality assessment for river mahanadi of odisha, India using statistical techniques and artificial neural networks, J. Clean. Prod. 417 (2023/09/10/2023) 137713.
[118] E.A. Freeman, G.G. Moisen, A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa, Ecol. Model. 217 (2008/09/24/2008) 48–58.
[119] T. Wu, S. Wang, B. Su, H. Wu, G. Wang, Understanding the water quality change of the Yilong Lake based on comprehensive assessment methods, Ecol. Indicat. 126 (2021/07/01/2021) 107714.
[120] E. Yongo, E. Mutethya, P. Zhang, S. Lek, Q. Fu, Z. Guo, Comparing the performance of the water quality index and phytoplankton index of biotic integrity in assessing the ecological status of three urban rivers in Haikou City, China, Ecol. Indicat. 157 (2023) 111286.
[121] O. Kisi, A.H. Dailr, M. Cimen, J. Shiri, Suspended sediment modeling using genetic programming and soft computing techniques, J. Hydrol. 450–451 (2012/07/11/2012) 48–58.
[122] C. Sang, L. Tan, Q. Cai, L. Ye, Long-term (2003-2021) evolution trend of water quality in the Three Gorges Reservoir: an evaluation based on an enhanced water quality index, Sci. Total Environ. 915 (2024).
[123] M.A.T. Koçer, H. Sevgili, Parameters selection for water quality index in the assessment of the environmental impacts of land-based trout farms, Ecol. Indicat. 36 (2014/01/01/2014) 672–681.
[124] W. Yuan, Q. Liu, S. Song, Y. Lu, S. Yang, Z. Fang, et al., A climate-water quality assessment framework for quantifying the contributions of climate change and human activities to water quality variations, J. Environ. Manag. 333 (May 1 2023).
[125] C. Zhang, X.Z. Nong, D.G. Shao, L.H. Chen, An integrated risk assessment framework using information theory-based coupling methods for basin-scale water quality management: a case study in the Danjiangkou Reservoir Basin, China, Sci. Total Environ. 884 (2023 Aug 2023).
[126] Nantong Water Resources Bulletin, Water Resources Bureau in Nantong, 2021.
[127] Nantong Water Resources Bulletin, Water Resources Bureau in Nantong, 2022.
[128] Yancheng Water Resources Bulletin, Water Resources Bureau in Yancheng, 2021.
[129] Yancheng Water Resources Bulletin, Water Resources Bureau in Yancheng, 2022.
[130] A. Chabuk, Q. Al-Madhlom, A. Al-Maliki, N. Al-Ansari, H.M. Hussain, J. Laue, Water quality assessment along Tigris River (Iraq) using water quality index (WQI) and GIS software, Arabian J. Geosci. 13 (Jul 13 2020).
[131] Q. L. Wang, Z. J. Li, J. N. Cai, M. S. Zhang, Z. D. Liu, Y. Xu, et al., "Spatially adaptive machine learning models for predicting water quality in Hong Kong," J. Hydrol., vol. 622, 2023 2023.
[132] L. Chen, M. Yang, Y. Liu, L. Nan, Early warning and joint regulation of water quantity and quality in the daqing River Basin, Water 14 (2022) 3068.