Environmental Science and Ecotechnology 20 (2024) 100412

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Spatiotemporal drivers of urban water pollution: Assessment of 102 cities across the Yangtze River Basin

Yi-Lin Zhao [a], Han-Jun Sun [a], Xiao-Dan Wang [b], Jie Ding [a, **], Mei-Yun Lu [a], Ji-Wei Pang [b, c], Da-Peng Zhou [d], Ming Liang [d], Nan-Qi Ren [a], Shan-Shan Yang [a, *]

[a] State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, China
[b] China Energy Conservation and Environmental Protection Group, Beijing 100082, China
[c] China Energy Conservation and Environmental Protection Group, CECEP Digital Technology Co., Ltd., Beijing 100089, China
[d] China Railway Engineering Design and Consulting Group Co., Ltd., Beijing 100055, China

ARTICLE INFO

Article history:
Received 25 June 2023
Received in revised form
6 March 2024
Accepted 8 March 2024

Keywords:
Basin management
Primary indices
Urban risk factors
Yangtze river basin
Local conditions

ABSTRACT

Effective management of large basins necessitates pinpointing the spatial and temporal drivers of primary index exceedances and urban risk factors, offering crucial insights for basin administrators. Yet, comprehensive examinations of multiple pollutants within the Yangtze River Basin remain scarce. Here we introduce a pollution inventory for urban clusters surrounding the Yangtze River Basin, analyzing water quality data from 102 cities during 2018–2019. We assessed the exceedance rates for six pivotal indicators: dissolved oxygen (DO), ammonia nitrogen (NH_3–N), chemical oxygen demand (COD), biochemical oxygen demand (BOD), total phosphorus (TP), and the permanganate index (COD_{Mn}) for each city. Employing random forest regression and SHapley Additive exPlanations (SHAP) analyses, we identified the spatiotemporal factors influencing these key indicators. Our results highlight agricultural activities as the primary contributors to the exceedance of all six indicators, thus pinpointing them as the leading pollution source in the basin. Additionally, forest coverage, livestock farming, chemical and pharmaceutical sectors, along with meteorological elements like precipitation and temperature, significantly impacted various indicators' exceedances. Furthermore, we delineate five core urban risk components through principal component analysis, which are (1) anthropogenic and industrial activities, (2) agricultural practices and forest extent, (3) climatic variables, (4) livestock rearing, and (5) principal polluting sectors. The cities were subsequently evaluated and categorized based on these risk components, incorporating policy interventions and administrative performance within each region. The comprehensive analysis advocates for a customized strategy in addressing the discerned risk factors, especially for cities presenting elevated risk levels.

1. Introduction

In recent years, the imperative to manage aquatic environments effectively has gained prominence, underscored by its inclusion as “Clean Water and Sanitation" within the United Nations' 2030 Sustainable Development Goals (SDGs). This goal is dedicated to ensuring universal access to clean drinking water and adequate sanitation facilities, and promoting water resources' sustainable management. China also attaches great importance to managing aquatic environments and water safety. In 2015, the Central Political Bureau's Standing Committee approved the “Water Pollution Prevention and Control Action Plan” (WPPCAP) to effectively expand efforts to prevent and control water pollution, protect national water security, and promote sustainable development. According to the plan's stipulations, by 2020, over 70% of section water in the seven major river basins in China, including the Yangtze and Yellow Rivers, is expected to be rated class III or better. Furthermore, the plan requires the Yangtze River Delta and Pearl River Delta regions to remove inferior class V cross-sections.

Spanning three major economic zones in eastern, central, and western China, the Yangtze River Basin is the largest in China,

* Corresponding author.
** Corresponding author.
E-mail addresses: dingjie123@hit.edu.cn (J. Ding), shanshanyang@hit.edu.cn (S.-S. Yang).

https://doi.org/10.1016/j.ese.2024.100412

Index of notations and abbreviations

BOD	Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
COD_{Mn}	Permanganate Index
DO	Dissolved Oxygen
IECM	Improved Export Coefficient Model
KMO	Kaiser–Meyer–Olkin
MAE	Mean Absolute Error
max_depth	The Maximum Depth
N	Nitrogen
NH_3–N	Ammonia Nitrogen
NANI	Net Anthropogenic Nitrogen Input (Howarth et al., 1996)
NAPI	Net Anthropogenic Phosphorus Input (Russell et al., 2008)
n_estimators	Number of Decision Trees
P	Phosphorus
PCA	Principal Component Analysis
PTA	Partial Triadic Analysis
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
SDGs	The United Nations 2030 Sustainable Development Goals
SHAP	SHapley Additive exPlanations
TP	Total Phosphorus
WPPCAP	Water Pollution Prevention and Control Action Plan

containing 19 provinces and 102 cities (including municipalities and autonomous regions) (Fig. 1), covering a total area of 1.8 million km^2. This region is highly valuable and has great development potential, playing an important role in China's economic and social development. The Yangtze River Basin was well managed during the implementation of the WPPCAP. By the end of 2020, the basin had no inferior class V sections. Before the evaluation, the percentage of exceedance events across the basin decreased from 14.33% in 2018 to 11.41% in 2019. Among the 24 nationally regulated water quality indicators, there are six major indices in the Yangtze River Basin: dissolved oxygen (DO), ammonia nitrogen (NH_3–N), chemical oxygen demand (COD), biochemical oxygen demand (BOD), total phosphorus (TP), and permanganate index (COD_{Mn}). Despite the progress, an analysis of water quality in 102 cities across 15 provinces/municipalities within the basin during 2018 and 2019 revealed that 12.87% of monitoring results surpassed the established standards, with exceedances recorded in 67 cities (Fig. 1). Regarding basin management, it is important that the drivers of the spatiotemporal variations of the exceedance rates of indices are promptly identified to provide corresponding control policies.

Numerous studies have analyzed the drivers of spatiotemporal changes in aquatic indices worldwide. For example, Slimani et al. [1] used partial triadic analysis (PTA) to evaluate the drivers of water quality change in the Medjerda River Basin (northern Tunisia) and found that the concentrations of NH_4^+, PO_4^{3-}, COD, and BOD in river water were strongly correlated with polluted urban sites, and confirmed that there was a strong relationship between land use and water quality. Kuriqi et al. [2,3] analyzed ecological impact data from 33 countries in five regions to assess the impacts

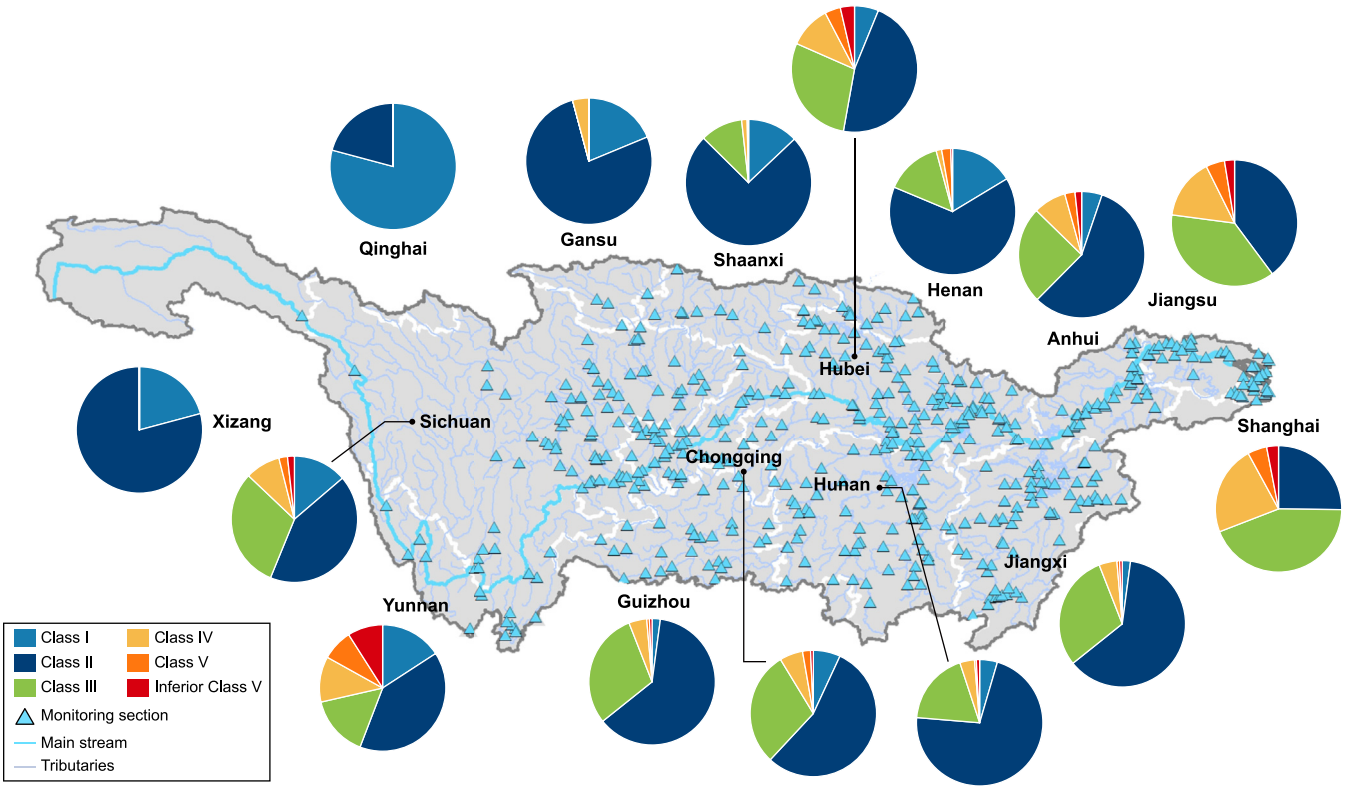

Fig. 1. Location of the 512 monitoring sections.

of small run-of-river hydropower plants on biota, water quality, hydrologic changes, and geomorphology, including altered flow regimes, reductions in bypassed stream reaches, and loss of longitudinal connectivity. They also estimated the impacts of nine hydrology-based environmental flow methods on hydropower production, altered flow regimes, and fish habitat conditions using hydropower, hydrological, and ecohydrological models. Cheng et al. [4] conducted a study on the Luanhe River Basin in northern China using an improved export coefficient model (IECM). They confirmed that the rural population, pigs, and arable land were the main contributors to TP concentrations. Using NANI (net anthropogenic nitrogen input), NAPI (net anthropogenic phosphorus input), and statistical models, Deng et al. [5] identified economic and land use factors as the main drivers of changes in anthropogenic N and P concentrations in the Yangtze River economic belt in China. However, previous studies have generally focused on small-scale watersheds or specific indices in large-scale watersheds. Furthermore, these methods mostly calculated pollutant loads based on the parameters provided by previous studies without comprehensively integrating the actual situation of each city or considering the environmental, industrial, and socioeconomic aspects holistically. Therefore, there is still room for further identification and analysis of primary indices in the Yangtze River Basin, and it is necessary to evaluate and analyze the drivers of spatiotemporal changes of primary indices based on the conditions of each city. Most existing research in China has focused on pollutant concentration changes, whereas relatively little attention has been paid to assessing water quality and adopting management methods. Therefore, further research is needed to align with China's water quality assessment and management system and provide references for managers.

Machine learning methods have been widely applied in environmental research in recent years. Random forest, a machine learning method that performs multivariate predictions well [6], has been extensively applied in recent research on water quality. Previous studies on water quality and pollution drivers in watersheds have used methods such as IECM (improved export coefficient model) and NANI [4,5], which calculate pollution loads based on known coefficients rather than analyzing pollution drivers based on water quality and urban attribute data from cities in the basin. Studies on machine learning have shown that random forest has strong advantages in terms of robustness and adaptive feature selection for this type of research [7,8] and can analyze the major pollution drivers based on the water quality and urban attribute data used as input. Few innovative studies have used the random forest model to analyze the comprehensive pollution drivers of multiple pollutants in large watersheds, such as the Yangtze River Basin. The random forest model can be divided into classification and regression. Random forest classification is mostly used for binary classification problems, such as identifying violations and risk warnings. For example, Scanlon et al. [8] determined the drivers of the spatiotemporal variability of drinking water quality in the USA using random forest classification. They found that arsenic and radionuclide violations were primarily related to semi-arid climates, whereas disinfection byproduct rule violations were primarily related to system operations. Kumar et al. [9] used random forest classification to assess the relationships between arsenic contamination of groundwater and parameters such as digital elevation model (DEM), land cover, and subsoil organic matter content in Jharkhand, India. Conversely, binary classification considers only two situations: exceeding or not exceeding the standard. This may group cases where contamination occurs only once a year with frequent contamination in the same category, thereby reducing the effective utilization of data. Random forest regression models are commonly used for predicting pollutant concentrations. For example, Li et al. [10] used random forest regression to predict the concentration of *Escherichia coli* on beaches in Lake Erie, USA; they identified water turbidity as the most important predictive factor, while accurate local wave height and rainfall data played a key role in model development. Khiavi et al. [11] used methods including random forest regression to create groundwater quality maps. However, owing to the multistage rating system used for water quality assessment in China, a concentration increase does not necessarily mean the standard is exceeded. Therefore, modeling based solely on concentration lacks intuitive interpretability for managers while requiring considerable computation. Considering the limitations of the existing methods, in this study, we conducted a random forest regression analysis on the exceedance rates of indices in urban sections. We used the advanced tree model interpretation tool SHapley Additive exPlanations (SHAP) to explain the model. This allowed us to analyze the spatiotemporal drivers of the occurrence probability of each index's exceedance events. This method retains information on the number of exceedance events and conforms to watershed management practices in the study area, thereby providing greater interpretability and accuracy.

Multivariate statistical analyses are also widely used in environmental research. For watershed studies, principal component analysis (PCA) is one of the most commonly used and developed multivariate statistical methods [12]. Daou et al. [13] used PCA to evaluate the spatiotemporal water quality patterns in four major rivers in southern, central, and northern Lebanon and the Bekaa Valley. They found that each river had different levels of eutrophication and pollution sources. Yang et al. [14] also used PCA to evaluate the drivers of spatiotemporal changes in surface water quality in the Xin'anjiang watershed, China and found that agricultural activities, erosion, and household and industrial emissions were the sources of water pollution in the region. However, most existing studies stop at data dimensionality reduction and risk factor identification and conduct little in-depth exploration of the scores of each case on each principal component. In this way, they fail to fully utilize the advantages of PCA. Therefore, in this study, we used PCA to identify urban risk factors and assess the risk levels of 102 cities in the Yangtze River Basin based on their scores for various risk factors, thereby fully leveraging the benefits of PCA and providing more intuitive guidance for watershed managers.

The main objectives of this study were to (1) construct a pollution inventory database for the urban agglomeration in the Yangtze River Basin and calculate the water quality exceedance rate for each of the 102 cities in the Yangtze River Basin from 2018 to 2019; (2) use random forest regression and SHAP to evaluate the main driving factors of the exceedances of each index; (3) use PCA to assess the risk factors for each city, and score and classify the cities in the basin based on the risk factors for the appropriate management; (4) propose management recommendations according to the characteristics of different cities based on the results. The novelty of this study is the comprehensive assessment of multiple major water quality indices in the Yangtze River Basin during 2018–2019 rather than just an analysis of individual contaminants. The study also considered various factors (26 in total), including environmental, socioeconomical, and industrial structure factors. Using advanced machine learning and statistical analysis, a method was proposed to analyze the drivers of water quality index exceedances in the Yangtze River Basin, identify urban risk factors, and present the results intuitively and clearly. Appropriate planning and management suggestions are provided, and the feasibility of the conclusions is validated by examining management measures and their effectiveness in the region over the past two years, thereby providing a reference for watershed managers (Fig. 2). The WPPCAP completed its round of acceptance in 2020, and the

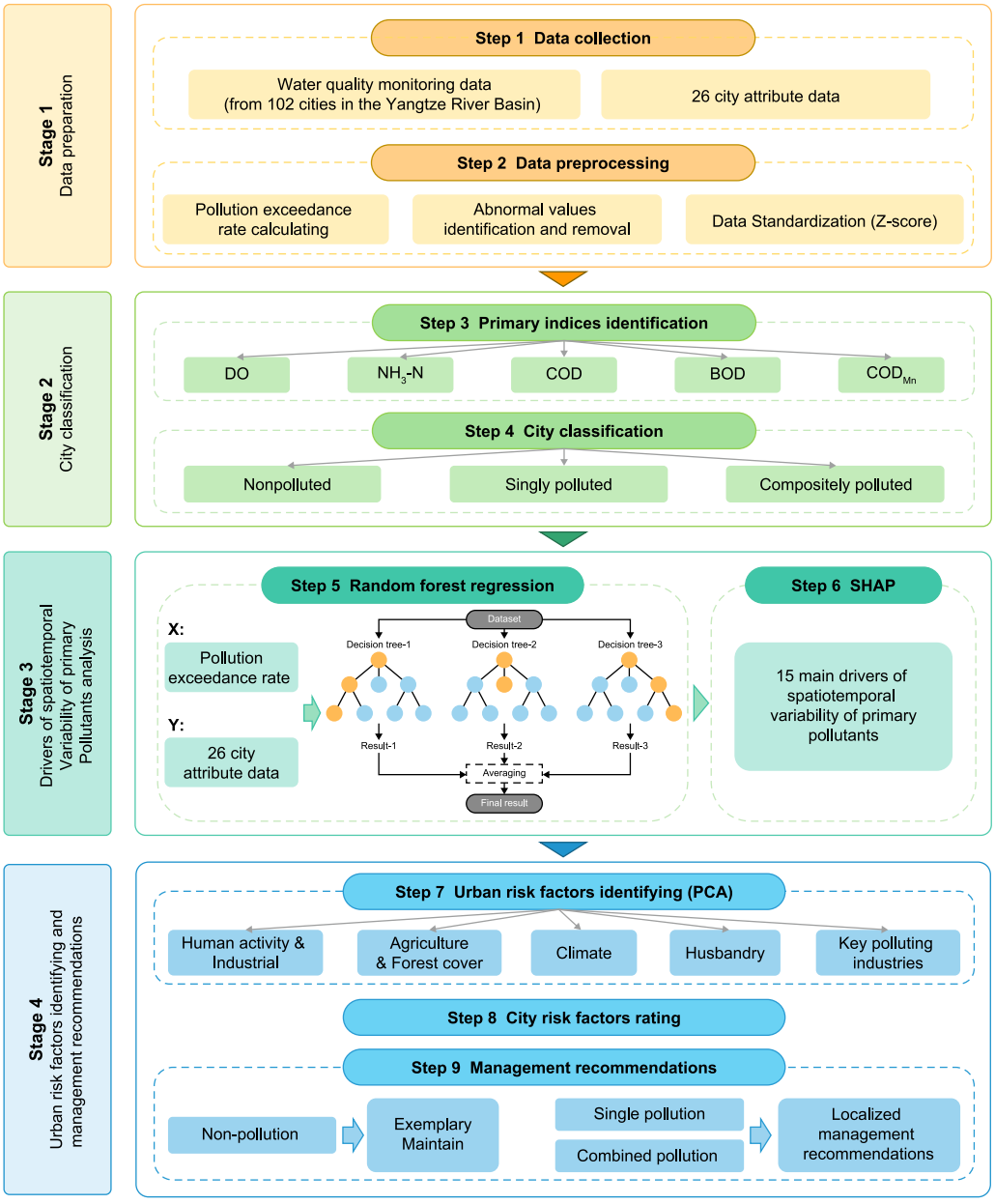

Fig. 2. Flowchart describing primary data and methods applied.

Ministry of Ecology and Environment and five other departments jointly issued the "Water Ecological Environment Protection Plan for Key Basins" in April 2023; this put forward the control objectives of the Yangtze River Basin for 2025. Hence, this study is significant as it provides decision-making references for managers in the Yangtze River Basin. This study reviews and reflects on the effect of water quality improvement in the Yangtze River Basin during the WPPCAP period and provides ideas and references for the future governance of the Yangtze River Basin.

2. Materials and Methods

2.1. Data sources

2.1.1. Water quality data

In China, river basin management is usually based on provinces and municipalities, whereas environmental, socioeconomic, and other data, such as urban attributes, usually come from annual statistical yearbooks. To give managers a clearer view and greater control of water quality pollution in the Yangtze River Basin, this study used annual water quality exceedance rate data for various indices in 102 cities within the basin for 2018 and 2019. Data in this study comprised monthly monitoring data from 512 sections (see Fig. 1 for section locations), including longitude, latitude, province, and city of the section, and concentrations of 24 indicators, including pH, DO, NH_3—N, COD, BOD, TP, permanganate index, fluoride, and arsenic. In total, 12,180 valid data entries were obtained. Following the current Surface Water Environmental Quality Standards in China, the water quality assessment uses a single-factor evaluation method, where each indicator is rated from class I to inferior class V based on its concentration. The water quality rating for each section is determined based on the lowest

rating of the indicators. Water quality is considered according to the standard when rated as class I, II, or III and exceeds the standard when rated as class IV, V, or inferior class V. A total of 1568 entries (12.87%) exceeded water quality standards. For any given index i in city k, the annual exceedance rate is calculated as follows:

$$P_{ki} = \frac{A_{ki}}{N_k \times 12 - A_{k-\text{null}}}$$

where P_{ki} represents the annual exceedance rate of index i in city k, A_{ki} represents the total number of exceedances of index i detected in all monitoring sections in city k for that year, N_k represents the total number of monitored sections in the river basin of city k. $A_{k\text{-null}}$ represents the total number of invalid monitoring data entries for all monitoring sections in city k for that year.

2.1.2. City attribute data

Through data review and literature research, we identified the factors that may cause the exceedances of the six major indices in the Yangtze River Basin from various dimensions, such as environment, population/socioeconomic, municipal/energy, agriculture, and industry. After data cleaning and preliminary modeling analysis, we selected 26 indicators as input for the random forest model (see Supplementary Material). Meteorological data, including temperature, atmospheric pressure, relative humidity, and precipitation, were obtained from monthly statistics provided by the National Basic Meteorological Station. Land use data were derived from the third national land survey results of provincial and municipal governments in 2019, China's latest land use data. Other city attribute-related data were obtained from provincial and municipal statistical yearbooks: the China Urban Statistical Yearbook, China Urban Construction Statistical Yearbook, and agricultural and industrial-related statistical yearbooks of various cities. All data were converted according to the land area of the city and were normalized using Z scores. Information on each indicator is presented in Table 1.

2.2. Methods

2.2.1. Random forest regression

The random forest regression model is an ensemble model based on decision trees; it is widely used in environmental risk assessment [15], pollutant concentration prediction [16], and other areas. The model combines multiple decision trees, each created using a randomly selected subset of the input variables. The final result is the average of all tree results [17].

In this study, 26 urban attribute data points from 102 cities in the Yangtze River Basin were used as the input feature vector x, and the annual exceedance rate of primary indices in the cities was used as the output variable. The original dataset was divided into training and validation sets (80% and 20% of the original dataset, respectively) to train the random forest regression model. Modeling was performed separately for the six primary indices DO, NH_3—N, COD, BOD, TP, and COD_{Mn}. Ten-fold cross-validation was used to ensure the accuracy of the results. Model performance was evaluated using three parameters: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) [16]. The parameters of each model, such as the number of decision trees (n_estimators) and maximum depth (max_depth), were adjusted to achieve optimal performance. All operations were performed using the scikit-learn package in Python 3.7.

2.2.2. SHapley Additive exPlanations (SHAP)

Although the random forest regression model exhibits good regression performance, it cannot explain the contribution of each feature to the prediction. Because this study aimed to identify the driving factors of primary index exceedances in cities to further evaluate urban risks and propose control measures, a reliable method was needed to explain the contribution of each feature and select those with the highest contributions as the main driving factors of primary index exceedances.

One of the most advanced analytical tools for tree models in recent years, namely SHAP, is based on the Shapley value. By calculating the average marginal contribution of one player in all possible combinations of other players in cooperative game theory

Table 1
Comparison of the names and contents of indicators.

Code	Potential sources of pollution	Unit	Abbreviation
PS01	Annual average air pressure	100 Pa	CLIM_AirPressure
PS02	Average annual temperature	°C	CLIM_Temperature
PS03	Annual average relative humidity	%	CLIM_Humidity
PS04	Annual precipitation	Mm	CLIM_Precipitation
PS05	Forest coverage rate	%	Forest_cover
PS06	Proportion of cultivated land area	%	LU_Plough
PS07	Proportion of urban, village and industrial and mining land area	%	LU_Urban&Miners
PS08	Proportion of land area for transportation	%	LU_Traffic
PS09	Proportion of land area for wetland, water and water conservancy facilities	%	LU_Water&Wetland
PS10	Population density	# km^{-2}	DESO_PopulDensity
PS11	Sewage Discharge	10000 m^3 km^{-2}	MUEN_Sewage
PS12	Amount of harmless treatment of domestic waste	t km^{-2}	MUEN_GarbageTreat
PS13	Amount of chemical fertilizer application	t ha^{-2}	AGRI_Fertilizer
PS14	Amount of pesticide use	t ha^{-2}	AGRI_Pesticides
PS15	Irrigated area	%	AGRI_Irrigation
PS16	Stockpile of pigs at the end of the year	# km^{-2}	AGRI_PigStock
PS17	Stockpile of sheep at the end of the year	# km^{-2}	AGRI_SheepStock
PS18	Annual poultry slaughter	# km^{-2}	AGRI_PoultrySold
PS19	Annual aquatic products production	t km^{-2}	AGRI_Aquatic
PS20	Total industrial assets	¥10000 km^{-2}	INDU_IndustrialAsset
PS21	Petroleum processing, coking and nuclear fuel processing industry assets	%	INDU_Petroleum
PS22	Chemical raw materials and chemical products manufacturing assets	%	INDU_Chemical
PS23	Non-ferrous metal assets	%	INDU_Nonferrous
PS24	Textile assets	%	INDU_Textile
PS25	Pharmaceutical manufacturing assets	%	INDU_Pharmaceutical
PS26	Leather, fur, feather and feather products and footwear industry assets	%	INDU_Leather

and allocating the absolute change in probability attributed to each explanatory variable, SHAP comprehensively considers the interaction between features and provides a more reliable estimate of feature importance [18]. The importance of each feature is given in the form of SHAP values. Therefore, the SHAP value calculation in this study was performed for the random forest regression models of the six primary indices, and the feature importance ranking for the 26 city attribute variables used as model input was provided. The top-ranked features in each model were selected as the driving factors of each primary index exceeding the standard, thereby more accurately evaluating the contribution of each driving factor to the index exceedance rate. All operations were performed using the SHAP package in Python 3.7.

2.2.3. Principal component analysis (PCA)

PCA is a data-dimensionality reduction technique. The main aim of PCA is to identify the underlying patterns and relationships between a set of observed variables and represent this information using a smaller number of uncorrelated principal component variables [19]. These components are linear combinations of the original variables, and each component captures a certain amount of data variation. The first principal component captures the maximum variation, and each subsequent component captures less variation. The Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test are commonly used to evaluate whether an original dataset suits PCA. The KMO measures the degree of common variance among the observed variables: a high KMO value (>0.5) indicates that the data are relatively compact and suitable for PCA. Bartlett's test checks whether there is a significant correlation among the observed variables: if the *p*-value is below a certain significance level (usually 0.05), then the null hypothesis (i.e., the assumption that the variables are uncorrelated) is rejected, and PCA can be applied. In this study, the KMO value was 0.735, and the significance level was less than 0.05 (Table 2), indicating that the dataset was suitable for PCA. The main factors identified by the random forest model for each index were used as input variables for the PCA, and the resulting principal components were used to identify the potential risk factors for each city. By analyzing the scores of each city for each principal component, this study provides guidance to watershed managers for identifying and controlling potential risks.

2.3. Limitations

The socioeconomic, agricultural, and industrial data used in this study were obtained from various statistical yearbooks published annually. To match the temporal scale of these data, we converted the water quality data into annual exceedance rates for each city. This may have reduced the precision of our study at the temporal scale. However, the richness of the study at the spatial scale compensates for the shortcomings at the temporal scale. As the Yangtze River Basin is more than 6000 km long and contains more than 100 cities, it is characterized by a large spatial span and high spatial variability among different cities. Considering that the main purpose of this study was to propose control recommendations based on the characteristics of different cities and provide a reference for city managers in the watershed, the importance of variability at the spatial scale is higher than that at the temporal scale; thus, the credibility of the study is not affected.

In addition, to match China's water quality monitoring and management model, we categorized only the cross-sectional water quality data into exceeding and not exceeding, which may have led to certain cities with high pollutant concentrations and others with relatively low pollutant concentrations being grouped. However, considering that the overall water quality of the Yangtze River Basin is good, with only 2% of inferior class V cross-sections, there are very few cities with extremely high pollutant concentrations; hence, the overall credibility of the study is not affected.

Table 2
KMO (Kaiser-Meyer-Olkin) and Bartlett's test.

KMO and Bartlett's test		
Kaiser-Meyer-Olkin measure of sampling adequacy		0.756
Bartlett's test of sphericity	Approx. Chi-Square	1904.972
	df	105
	Sig.	0.000

3. Results and discussion

3.1. Water quality profile of the Yangtze River Basin urban agglomeration and classification of cities based on exceedance of the standard

The exceedances of the six primary indices in each city of the Yangtze River Basin urban agglomeration were counted, and a heat map of the exceedance rates of the primary indices in cities was created (Fig. 3). Overall, areas with severe pollution include Shanghai and some cities in Anhui (such as Chuzhou, Xuancheng, etc.) and Hubei (such as Qianjiang, Tianmen, Wuhan, etc.) provinces. Shanghai exceeded the standards for all six indices, with the DO exceedance rate reaching 22.73% in 2018, whereas Chuzhou (Anhui Province) had a serious COD exceedance rate of 55.56% in 2019. Regions with relatively good water quality include the water source areas of Xizang Zizhiqu, Qinghai Province, and certain cities in Shaanxi Province. Notably, Changdu (Xizang Zizhiqu) and Yushu (Qinghai Province), located in the upper reaches, did not show any exceedances, reflecting good water quality at the source of the Yangtze River. Based on the pollution levels in cities within the Yangtze River Basin urban agglomeration, the cities can be classified into three categories for management purposes.

(1) Non-polluted: Cities that did not exceed the standard for any of the six indices, such as Changdu (Xizang Zizhiqu) and Yushu (Qinghai Province) (Fig. 3c). During 2018 and 2019, non-polluted cities accounted for 20.6% of all cities in the basin. This type of city accounted for 30.4% of all cities in 2018 and 25.5% in 2019, indicating that water quality conditions in the basin improved over the study period.
(2) Singly polluted: Cities with only one index exceeding the standard, such as Leshan in Sichuan Province (TP exceedance rate of 4.202%) and Xiantao in Hubei Province (COD exceedance rate of 4.165%) (Fig. 3c). Singly polluted cities accounted for 13.7% of all cities in the basin between 2018 and 2019. This type of city accounted for 7.8% and 22.5% of all cities in 2018 and 2019, respectively. In 2018, the singly polluted cities had the highest NH_3–N exceedance rates (62.5%); in 2019, TP exceedance rates were the highest (39.1%). This suggests that the problems associated with TP and NH_3–N should be emphasized.
(3) Compositely polluted: Cities with two or more indices exceeding the standard, such as Shanghai (with all six indices exceeding the standard) and Nanjing in Jiangsu Province (with all six indices exceeding the standard) (Fig. 3c). From 2018 to 2019, compositely polluted cities accounted for 65.7% of all cities in the basin. This type of city accounted for 61.8% of all cities in 2018 and 52% in 2019. A total of 40.3% of compositely polluted cities had all six indices exceeding the standard during 2018–2019. Overall, the data show that the pollution situation of cities in the Yangtze River Basin is

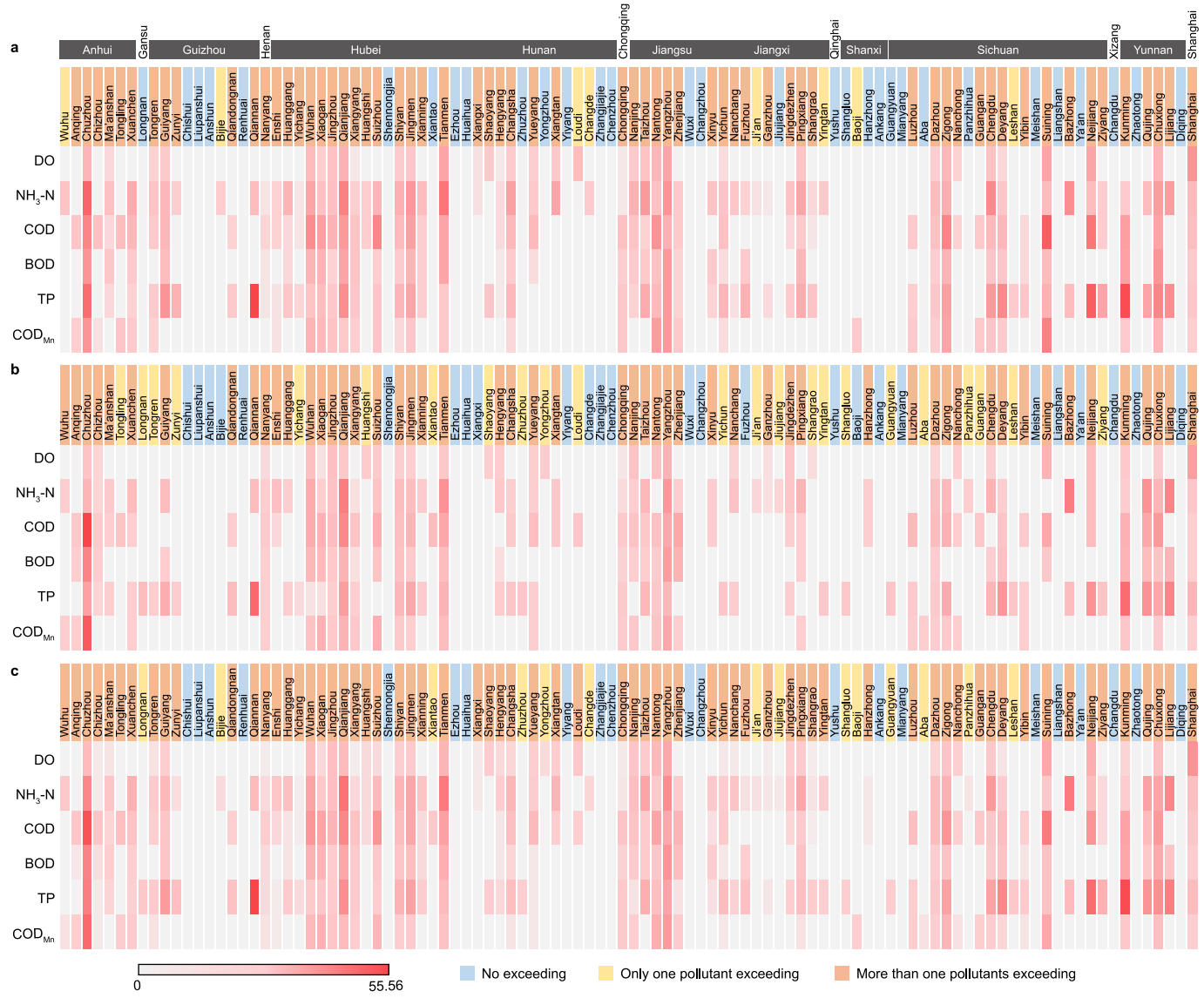

Fig. 3. Heat map depicting the exceedance of primary indices in cities within the Yangtze River Basin urban agglomeration. The intensity of the red squares in the figure represents the degree of exceedance rate, while the dark gray blocks indicate the provinces where the cities are located. The different background colors of city names indicate different city types, with blue representing non-polluted cities, yellow representing single-polluted cities, and orange representing composite-polluted cities. **a**, Exceedance situation in 2018. **b**, Exceedance situation in 2019. **c**, Exceedance situation for the two years of 2018 and 2019.

complex, and it is necessary to propose targeted control recommendations according to the pollution situation of different cities.

Cities that do not exceed the pollution standard should maintain strict pollution control measures, implement existing policies, and propose appropriate new environmental governance measures. Furthermore, it is important to encourage other provinces and cities to use these successful practices as examples. In singly and compositely polluted cities, further analysis is required to identify the drivers causing index exceedances and risk factors associated with pollution in each city. Accordingly, tailored governance measures should be proposed based on local conditions. Compositely polluted cities can improve their status by implementing appropriate measures to become singly polluted or non-polluted. Similarly, singly polluted cities can eliminate pollution through effective governance measures. For example, in 2018, Fuzhou (Jiangxi Province) experienced exceedances in NH_3—N (2.08%), BOD, and TP (12.5%) (Fig. 3a) and was therefore compositely polluted; however, after implementing appropriate governance measures, it transitioned into a non-polluted city in 2019 (Fig. 3b).

3.2. Analysis of spatiotemporal drivers of primary indices in the Yangtze River Basin urban agglomeration

To identify the spatiotemporal drivers of primary indices in the Yangtze River Basin urban agglomeration and propose targeted control recommendations, this study utilized a model that incorporates city attribute data normalized using *Z* scores with the DO, NH_3—N, COD, BOD, TP, and COD_{Mn} exceedance rate data. After comparing the performances of the different models, we selected random forest as the analysis tool for this study. Random forest regression models were constructed for the annual exceedance rate of each model, and reliable models were obtained after parameter adjustment (see Supplementary Material for model result

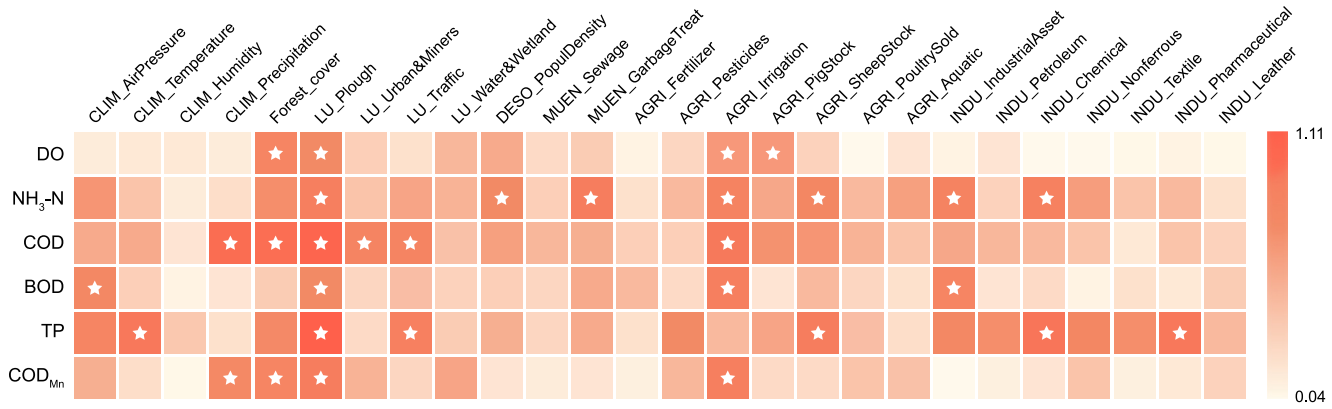

Fig. 4. Heatmap of the primary index spatiotemporal drivers in the Yangtze River Basin obtained through SHAP random forest regression model analysis. In the heatmap, darker color blocks indicate higher SHAP values, indicating a greater contribution of the corresponding factor to index exceedances. Factors marked with a pentagram symbol represent the top-ranking factors with high SHAP values among the driving factors for each primary index.

parameters). The SHAP method was employed to interpret the models and identify the importance of spatiotemporal drivers for the extent of primary index exceedance in the Yangtze River Basin.

Among the identified spatiotemporal driving factors, cropland coverage and irrigation area contributed significantly to the exceedance rates of all six indices, reflecting the impact of agricultural activities on water quality (Fig. 4). Cropland coverage had particularly high SHAP values of 1.03 and 1.11 for COD and TP, respectively. Forest cover was the most important factor leading to insufficient DO and significantly influenced the NH_3—N, COD, and COD_{Mn} exceedance rates, with a SHAP value of 0.86 in the COD model. Industrial factors also affected multiple indices. The proportion of industries was an important factor in increased NH_3—N and BOD exceedance rates. Among the key industries mentioned in the WPPCAP, pharmaceutical manufacturing was a significant contributor to TP exceedance, while raw chemical materials and chemical product manufacturing had a considerable influence on the NH_3—N and TP exceedances, with a SHAP value of 0.74 in the TP model. In the livestock industry, sheep stocking was an important cause of the NH_3—N and TP exceedances, whereas pig stocking strongly contributed to insufficient DO. Among human-related factors, the amount of municipal solid waste disposal and population density significantly impacted the NH_3—N exceedance. The urban, rural, and mining land areas had considerable impacts on COD exceedance, whereas the transportation land area was an important factor for the TP exceedance, with a SHAP value of 0.54. Regarding climate-related factors, precipitation and atmospheric pressure strongly influenced the COD and BOD exceedances, with precipitation having a high SHAP value of 0.89 in the COD model. In contrast, temperature had a greater impact on the TP exceedance, with a SHAP value of 0.68 in the model.

To further determine the relationships between the high-ranking SHAP values of the driving factors and the exceedance rates of each index, we examined the correlations between the top-ranked spatiotemporal driving factors and the exceedance rates of various indices (Fig. 5). The driving factors are sorted based on their importance, with the importance decreasing from top to bottom. The horizontal (x) axis represents the SHAP values: SHAP values < 0 indicate a negative contribution; SHAP values $= 0$ indicate no contribution; and SHAP values > 0 indicate a positive contribution. Positive contributions indicate that these features increase the index exceedance rate, whereas negative contributions indicate the opposite effect. The colors indicate whether the driving factor has a high (red) or low (blue) influence on the exceedance rate prediction [18]. Cropland coverage and irrigation area were strongly positively correlated with the exceedance rates of various indices. In particular, cropland coverage had the highest SHAP value (>8.0) for predicting the NH_3—N, COD, and COD_{Mn} exceedance rates. Studies have shown that agricultural non-point source pollution contributes significantly to water pollution in China [20,21]. Xu et al. [22] used a conceptual model for drainage basin water quality and non-point pollution to study the effects of agricultural fertilization practices on NH_3—N concentrations in 18 cities in the Yangtze and Yellow River Basins. The results showed that agricultural behavior has a strong positive effect on NH_3—N pollution. Duan et al. [23] established a national non-point source pollution database and estimated the N and P nutrient loads from county-level crop cultivation in 2015; they found that nutrient surpluses were relatively high in areas south of the Yangtze River. Cui et al. [24] studied the input and distribution characteristics of anthropogenic P in the Yangtze River Basin and sub-basins and analyzed the driving factors; they found that the major anthropogenic P inputs to the middle and lower Yangtze River plains were from agricultural sources. Previous studies on the impacts of agriculture on water pollution have focused on nutrients such as N and P, with less discussion on indices such as DO and COD, while the study areas have mostly been parts of the Yangtze River Basin. By analyzing six major water quality indices in 102 cities in the Yangtze River Basin, our study shows that agricultural factors have a greater impact on all water quality indicators in the Yangtze River Basin. The production and use of fertilizers and pesticides generate a high volume of reducing substances — mainly organic pollutants. Irrigation and other activities cause the flow of these nutrients into water bodies, increasing the levels of reducing organic substances and leading to COD, BOD, and COD_{Mn} exceedances [25]. Sewage and feces from agricultural activities and using fertilizers and pesticides can contribute to NH_3—N and P pollution in water bodies, potentially causing eutrophication [26]. The decomposition of organic matter in water consumes DO, which decreases the DO content in aquatic environments [27]. As non-point source pollution in the basin is receiving increasing attention, managers should pay more attention to agricultural surface pollution and introduce timely relevant policies to control the flow of agricultural pollutants into water bodies. Forest cover is negatively correlated with the exceedance rates of various indices, with the minimum SHAP value approaching −2.0 in the COD model. Forests are beneficial for capturing nutrients, preventing soil erosion and loss, and reducing eutrophication, which can help reduce the possibility of NH_3—N, COD, and COD_{Mn} exceedances [28]. Additionally, forests generate large amounts of oxygen through photosynthesis, thereby

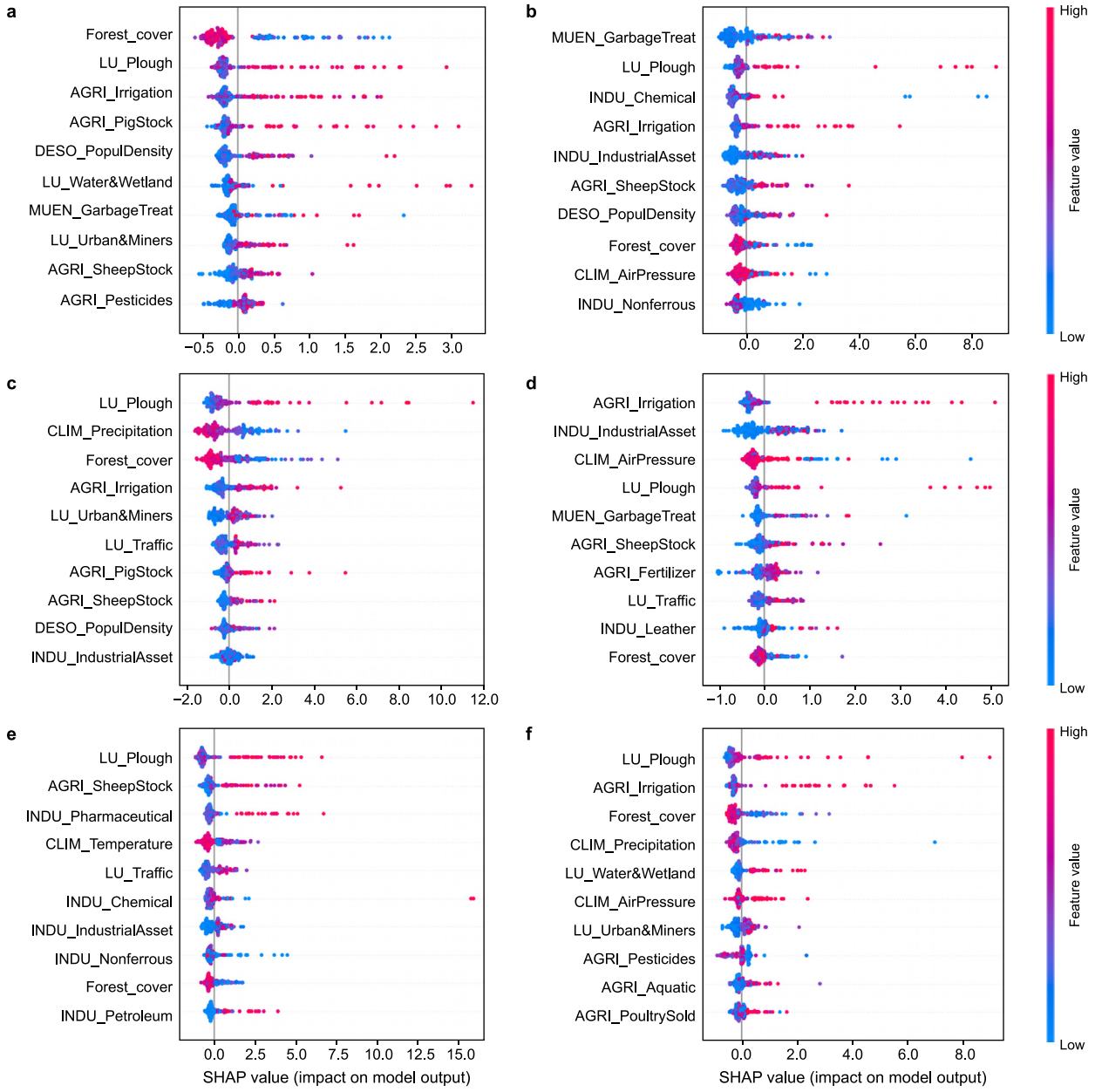

Fig. 5. Correlation between the top ten driving factors and exceedance rates of each primary index: **a**, DO; **b**, NH_3—N; **c**, COD; **d**, BOD; **e**, TP; **f**, COD_{Mn}. The red color represents higher values of the driving factors, while the blue color represents lower values. The *x*-axis origin indicates a positive impact on exceedance rates to the right and a negative impact to the left. Taking panel **a** as an example, forest coverage is negatively correlated with exceedance rates, while cropland area is positively correlated with exceedance rates.

promoting atmospheric reoxygenation processes and enhancing DO concentrations in water [29]. This indirectly improves the situation regarding the NH_3—N, COD, and COD_{Mn} exceedances. Watershed managers should realize the importance of forests for water quality improvement and increase afforestation activities.

The proportion of industries was positively correlated with the NH_3—N and BOD exceedance rates. Previous studies have discussed the impacts of industries on the Yangtze River Basin. A study on the water quality along the mainstem of the Yangtze River showed that industry is an important source of pollution in the Yangtze River Basin [30]. Peng et al. [31] analyzed the linkage between industrial production and water pollution and its drivers in the Yangtze River Basin provinces from 2012 to 2017. The results showed that the chemical industry was the main source of COD, NH_3—N, and TP emissions in the Anhui, Jiangsu, Jiangxi, and Hunan Provinces. Most previous studies have focused on the provincial scale or have considered industrial pollution sources as one component. Our study was conducted at a more precise (i.e., municipal) scale and analyzed the impact of different industrial sectors on water quality. Industrial wastewater from coking plants and synthetic ammonia fertilizer factories contains high concentrations of NH_3—N [32,33]. Industrial pollutants are significant sources of BOD in water bodies [34]. The chemical raw material and manufacturing industry positively influences the NH_3—N and TP exceedances. This industry includes N- and P-containing fertilizer manufacturing, pesticide manufacturing, and other industries prone to N and P pollution [35]. However, the pharmaceutical manufacturing industry significantly impacts the TP exceedance rate, with a maximum SHAP

value approaching 7.5, possibly due to the high P content of pharmaceutical wastewater [36]. Industrial pollution remains a non-negligible problem for aquatic environment management in the Yangtze River Basin, and managers should pay attention to the discharge of key industries, such as the chemical and pharmaceutical industries, to avoid industrial pollution aggravation. Regarding husbandry, pig stocking has a strong positive effect on the DO exceedance rate, whereas sheep stocking has a strong positive influence on the TP and NH_3—N exceedance rates, with the maximum SHAP value exceeding 5.0 in the TP model. This may be due to the discharge of nutrients, such as N and P, from feed and livestock manure into aquatic environments and their oxidation process consuming DO in the water [37–39]. Simultaneously, sheep may damage the soil during grazing, resulting in an easier influx of nutrients into water bodies and eventually causing eutrophication [40]. Previous studies have shown that the negative impacts of livestock farming on water quality pollution may accumulate over time [41]; hence, managers should optimize the regional layout of livestock and poultry farming, introduce advanced feeding techniques, and vigorously promote the resourceful use of waste from livestock and poultry farming [42]. Regarding other human-related factors, domestic waste disposal and population density were positively correlated with the NH_3—N exceedance. The transportation land area was positively correlated with the COD and TP exceedance rates, while the urban, rural, and mining land areas were also positively correlated with the COD exceedance. During landfilling and transporting domestic waste, leachate containing NH_3—N can enter the water and cause pollution [41]. Domestic wastewater, feces, and garbage from daily human activities are significant sources of NH_3—N, and a higher population density is more likely to result in NH_3—N exceedance [43]. Greater urban, rural, and mining land areas indicate increased human activity and the associated wastewater discharge from production and daily life may lead to COD exceedance [44]. Runoff from roads contains high concentrations of TP and reducible substances, which can cause TP and COD exceedances after entering aquatic environments [45]; therefore, a greater transportation land area undoubtedly increases the risk of exceedance. Cities with severe pollution in the Yangtze River Basin, such as Shanghai, Wuhan, and Chengdu (where all six indicators have exceeded the standard), had higher population densities, domestic waste disposal volumes, and transportation land areas. For these areas with frequent human activities, managers should closely monitor domestic pollution sources, introduce measures to limit the discharge of pollutants from domestic sources, such as garbage classification and vehicle traffic restrictions, and increase environmental protection publicity to raise the public's awareness of aquatic environment protection.

Among the climate-related factors, precipitation was negatively correlated with the COD exceedance, with a minimum SHAP value close to −2.0. This suggests that decreased precipitation reduces the river water volume, resulting in higher pollutant concentrations [46]. Conversely, atmospheric pressure was negatively correlated with the BOD exceedance. Many areas with BOD exceedance also experience DO exceedance, which may be because decreased atmospheric pressure leads to a reduction in the oxygen content of the water, thereby affecting the oxidation process of reducible substances and causing BOD exceedance [47]. The relationship between temperature and TP exceedance is not linear. This may be because certain P-removing microorganisms have a suitable range of temperatures for their survival and reaction, thereby affecting the TP concentration in water [48]. We found positive correlations between the wetland, water, and water conservancy facility areas and the exceedance rates of multiple indices, such as COD_{Mn} and DO. This finding contradicts our common understanding; however, it may be attributed to larger water body areas implying a higher number of monitoring sections in the respective cities. Consequently, the likelihood of exceedances increases, posing greater pollution control challenges and higher exceedance rates. These climate-related factors are attributes of the city itself and are difficult to change in the short term. This study serves as a reminder to managers that more attention is necessary regarding preventing pollution caused by these factors, especially during extreme weather (e.g., heavy rainfall and high temperatures), to minimize the damage caused by water pollution.

3.3. Assessment of urban risk factors and discussion of control recommendations

3.3.1. Urban risk factors analysis

After analyzing the primary index driving factors, it is necessary to propose tailored control measures based on the specific conditions of each city. However, the factors contributing to water quality exceedances in a city can be multifaceted; therefore, it is important to identify the risk factors specific to each city for effective control. We conducted a dimensionality reduction analysis using PCA on the top 15 spatiotemporal driving factors extracted from the SHAP analysis of the index exceedance rates to identify urban risk factors. When the number of principal components was set to five, 79.976% of the variance was explained. Therefore, we identified the following five principal components as urban risk factors (Table 3): (1) human activity and industrial factor, (2) agriculture and forest cover factor, (3) climate factor, (4) husbandry factor, and (5) key polluting industry factor. We calculated the scores of each city in the Yangtze River Basin for these five principal components. We examined their correlations with the probability of exceeding the pollution threshold (exceedance of at least one index). Except for principal component 3 (climate factor), which had a negative correlation with the exceedance probability, the scores of the remaining principal components were positively correlated with the exceedance rates of cities (Table 4). The specific characteristics of each risk factor are as follows.

Human activities and industrial factors. The human activity and industrial factor explains 28.60% of the overall variance and has higher loads on several characteristics of urban, rural, and mining land, transportation area, population density, domestic waste disposal, and industrial assets. This indicates that it mainly results from the impacts of human and industrial activities. Cities with high scores for this risk factor typically have high values for one or more related indicators. Representative cities include Shanghai, Wuhan (Hubei Province), and Nanjing (Jiangsu Province), where priority should be given to implementing policies and measures related to municipal and industrial activities that can effectively improve pollution indicators such as NH_3—N and BOD. For example, the “Implementation Plan for Urban and Rural Domestic Waste Treatment in Jiangsu Province” introduced at the end of 2018 in Jiangsu Province contains treatment measures such as leachate treatment and safe disposal of fly ash, which have reduced the overall NH_3—N exceedance rate by 5.54%.

Agriculture and forest cover factor. The agricultural and forest cover factor significantly influences arable land area, irrigated area, and forest cover rate, explaining 19.54% of the overall variance. Considering the significant negative correlations between the forest cover rate and each pollution exceedance rate, this risk factor primarily reflects the impacts of agricultural activities and insufficient forest cover. Cities that score high on this risk factor, such as Tianmen (Hubei Province) and certain cities in Jiangsu Province, typically have high proportions of agricultural activities and low forest cover. In these cities, priority should be given to

Table 3
Results of PCA.

Code		Groupings 1	2	3	4	5
Grouping 1: Human activity and industrial						
PS	LU_City_Mine	0.826	-	-	-	-
	LU_Traffic	0.777	-	-	-	-
	DESO_PopuDens	0.962	-	-	-	-
	MUEN_GarbageTreat	0.954	-	-	-	-
	INDU_Asset	0.960	-	-	-	-
Grouping 2: Agriculture and forest cover						
	Forest_cover	-	−0.830	-	-	-
	LU_Plough	-	0.909	-	-	-
	AGRI_Irrigation	-	0.863	-	-	-
Grouping 3: Climate						
	CLIM_PRS	-	-	0.708	-	-
	CLIM_TEM	-	-	0.853	-	-
	CLIM_PRE	-	-	0.690	-	-
Grouping 4: Husbandry						
	AGRI_PigStock	-	-	-	0.777	-
	AGRI_SheepStock	-	-	-	0.828	-
Grouping 5: Key polluting industries						
	INDU_Chemical	-	-	-	-	0.757
	INDU_Medical	-	-	-	-	0.784
Eigenvalue		4.290	2.931	1.978	1.495	1.303
Variance (%)		28.601	19.539	13.186	9.965	8.686
Cumulative variance (%)		28.601	48.140	61.326	71.291	79.976

Table 4
Correlation between the scores of each principal component and the probability of exceedance events in Yangtze River Basin cities.

Variable	Parameter	Grouping 1: Human activity and industrial	Grouping 2: Agriculture and forest cover	Grouping 3: Climate	Grouping 4: Husbandry	Grouping 5: Key polluting industries
Probability of exceeding	Pearson correlation	0.179*	0.320**	−0.276**	0.161	0.211*
	Sig.(2-tailed)	0.030	0.000	0.001	0.051	0.010
	number	147	147	147	147	147

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

implementing green agriculture-related policies to control agricultural non-point source pollution. Additionally, appropriate afforestation activities can enhance the forest cover. Agricultural factors and forest cover affect nearly all pollution indicators; therefore, managing this risk factor can effectively reduce pollution in the entire watershed. For example, in Hubei and Jiangsu provinces, where arable land and irrigated areas account for a significant proportion and where agricultural activities are frequent, green agriculture-related policies and control measures have been implemented. In October 2018, Hubei Province issued a notice to promote arable land quality protection and fertilizer reduction to improve efficiency. The same year, Jiangsu Province introduced an implementation judgment to accelerate green agricultural development. These measures reduced the index exceedance rates in most areas of these two provinces. Sichuan Province has consistently emphasized afforestation; in 2019, it completed afforestation of an area of 400,370 ha, significantly alleviating exceeding levels of indicators such as DO, COD, and COD_{Mn} in most parts of the province.

Climate factor. The climate factor explains 13.19% of the overall variance, with a strong emphasis on atmospheric pressure, annual precipitation, and annual average temperature, indicating its influence on pollution concerning climate and geographical characteristics. The climate factor scores are negatively correlated with the exceedance rates of all indices (Table 5), suggesting that cities with lower climate factor scores are more prone to exceedances. This may be because the geographical and climate factors are complex and comprehensive. Yunnan Province serves as a typical example of a region with a lower climate factor score: the exceedances in Yunnan Province indicate that it might be influenced by its unique climate and geographical factors. As natural attributes of a city, climate factors remind managers to pay more attention to the geographic characteristics of the city itself and prevent natural disasters (e.g., droughts, floods, and high temperatures) that may cause the prompt exceedance of water quality

Table 5
Correlation between the scores of principal component 3 (climate factor) and the exceedance rate of each pollutant.

Variable	Parameter	TP	DO	NH_3−N	COD	BOD	COD_{Mn}
Grouping 3: Climate	Pearson Correlation	−0.470[b]	−0.019	−0.214[a]	−0.181	−0.294[b]	−0.132
	Sig.(2-tailed)	0.000	0.836	0.021	0.052	0.001	0.157
	Number	116	116	116	116	116	116

[a] Correlation is significant at the 0.05 level (2-tailed).
[b] Correlation is significant at the 0.01 level (2-tailed).

indices. Simultaneously, after controlling other risk factors, an emphasis on climate factors can further improve the management effect. Therefore, planning and management strategies should be developed to consider regional climate characteristics when proposing targeted measures for improvement.

Husbandry factor. The husbandry factor explains 9.97% of the overall variance, with a higher load on pig and sheep farming, indicating its influence on pollution concerning the livestock industry factors. Cities with higher scores for this factor typically have more pig and/or sheep farming. For such cities, the implementation of husband-related policies should be prioritized. Based on the actual conditions of the region, optimization and regulation of the pig or sheep farming industries can effectively improve the exceedance of pollutants such as DO and TP. For example, in 2018, Jiangsu and Hubei provinces were affected by African swine fever, significantly impacting the pig farming industry. Jiangsu Province issued a response plan that included measures such as culling, the prohibition of farming, and restrictions. The reduced pig population decreased the excess DO in Jiangsu Province by 2.41%.

Key polluting industry factors. The key polluting industry factor explains 8.69% of the overall variance, with a strong load on pharmaceutical manufacturing and chemical raw materials and chemicals manufacturing, indicating their primary association with certain key polluting industries. Cities with higher scores for this factor, such as Guiyang, Qiannan Prefecture (Guizhou Province), and Kunming (Yunnan Province), have typically high proportions of these industries. For these cities, priority should be given to regulating these industries and implementing relevant policies to control water pollution. Effective control measures can significantly improve the pollution situation regarding NH_3—N and TP.

3.3.2. City risk factor ratings and control recommendations

After completing the PCA of the urban risk factors, the scores of each city for each risk factor were arranged in ascending order and divided equally into 10. Each risk factor for each city was then rated on a scale of 1—10; for example, a rating of 1 indicated that the city's score on that principal component was in the lowest 10% of all data, while a rating of 10 indicated that the city's score was in the top 10%. The scoring method for the climate factor was opposite to that for the other four risk factors, meaning that a higher score corresponded to a lower rating. By plotting radar charts for each city, managers can quickly identify the risk factors that may contribute to pollution exceedances in a specific area and initiate further investigations and corresponding control measures (see Table S1for the scores of all cities for each risk factor). Fig. 6 shows radar charts of the representative cities. Nantong (Jiangsu Province) scores high for all five risk factors, whereas Shanghai scores higher for human activities, industrial, and climate factors. Notably, some areas may have multiple risk factors with simultaneously high ratings, indicating that the factors causing pollution exceedances in these areas are more complex. Therefore, a comprehensive set of policies should be developed based on the actual situation, and initiating a macro-level action plan for aquatic environmental protection should be a priority. The evaluation of pollution control in the Yangtze River Basin from 2018 to 2019 (Fig. 7) shows that many areas improved their pollution exceedances, closely related to the policies implemented in these regions. However, the improvement in pollution exceedance rates remains inadequate in some areas and has even increased in others. For example, the orange area in Fig. 7c indicates an increase in the COD exceedance rate in cities such as Qujing (Yunnan Province), Chuzhou (Anhui Province), and Zhenjiang (Jiangsu Province). Regions with poor pollution improvement should use cities with effective pollution control

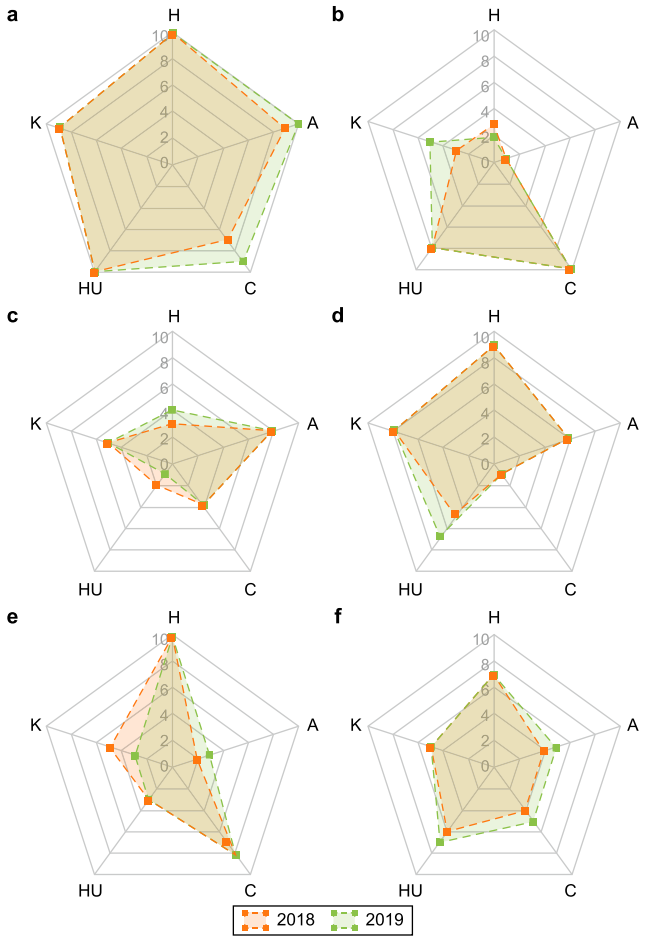

Fig. 6. Risk assessment results for representative cities in the Yangtze River Basin. **a**, Nantong; **b**, Chuxiong; **c**, Anqing; **d**, Nanchang; **e**, Shanghai; **f**, Chongqing. H: human activity and industrial; A: agriculture and forest cover; C: climate; HU: husbandry; K: key polluting industries.

measures as examples and implement relevant policies. For example, the Yunnan Province has relatively few agricultural and domestic waste management policy measures, resulting in poor NH_3—N control. Yunnan Province should follow the example of areas such as Shanghai and Jiangsu Province and implement appropriate control measures while considering its geographical environment.

Besides formulating control measures for highly rated risk factors, environmental protection policies and increasing efforts in water management have contributed to alleviating the overall water pollution in the basin. For example, as an area with a relatively high overall exceedance rate, Shanghai issued the “Shanghai Environmental Protection and Construction Three-Year Action Plan for 2018—2020” in March 2018, which strengthened environmental protection efforts, increased the intensity of aquatic environment management, and focused on agricultural pollution control. The strictest water resource management system was implemented in Jiangsu Province in 2018, with various departments coordinating their efforts to improve the aquatic environment. As a result, Shanghai and Jiangsu Province reduced their overall exceedance rates by 5.68% and 6.18%, respectively, from 2018 to 2019. Furthermore, targeted control plans for specific pollutants have also contributed to exceedance rate improvements, particularly in cities with a single pollution source. For instance, many areas of Guizhou Province experienced significant TP exceedances. In 2018,

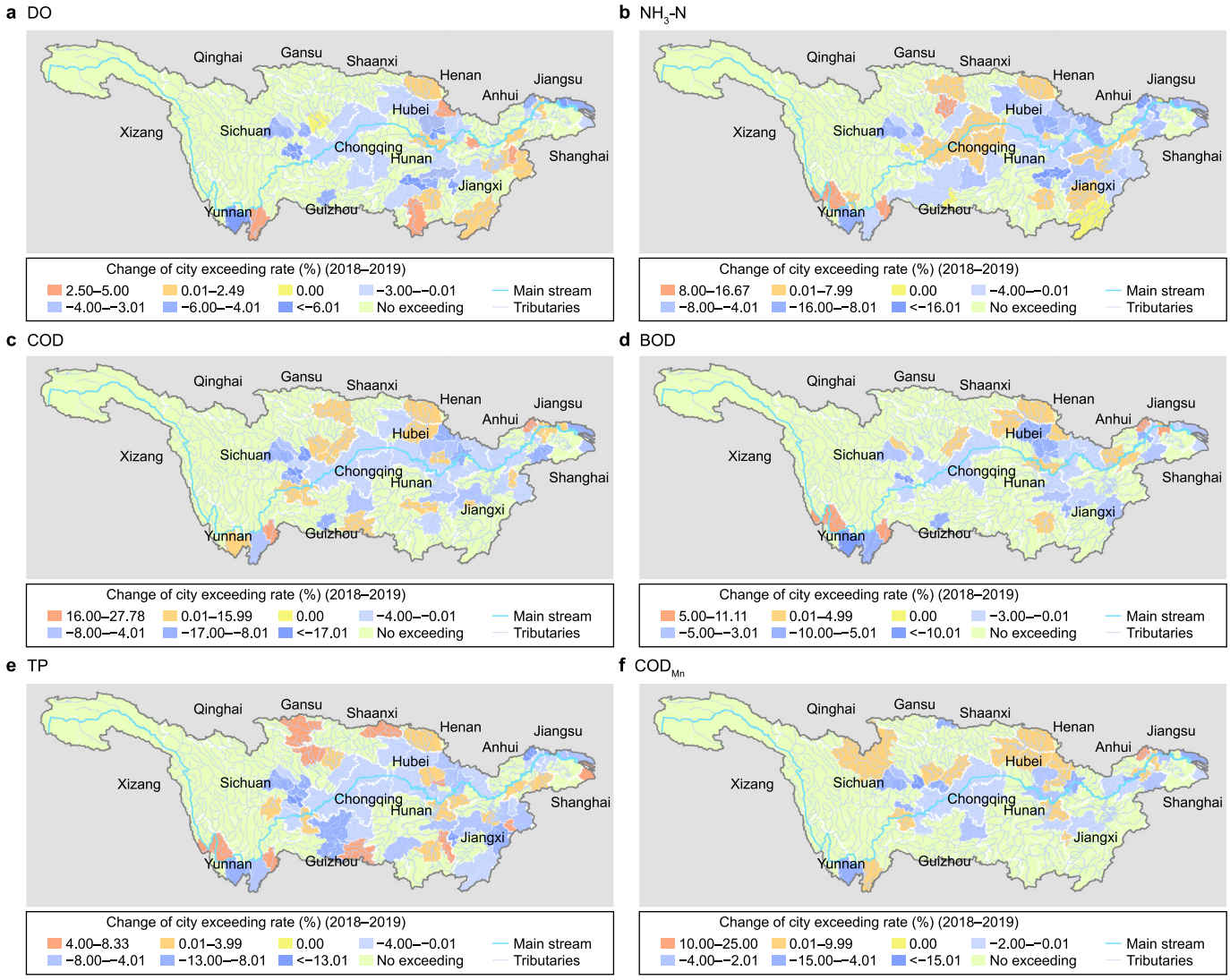

Fig. 7. Changes in index exceedance rates in the Yangtze River Basin from 2018 to 2019: **a**, DO; **b**, NH_3—N; **c**, COD; **d**, BOD; **e**, TP; **f**, COD_{Mn}. The color indicates the change in cities' exceedance rates, with blue representing a decrease in exceedance rate, orange representing an increase in exceedance rate, and yellow representing no change. The intensity of the color represents the magnitude of the change.

Guizhou Province issued a notice on accelerating the comprehensive utilization of P resources and reducing the TP exceedance rate by 4.44% across the province. Other areas that face exceedances of a single pollutant can use this approach as an example and implement targeted control measures for specific pollutants.

However, although the proportion of industries is an important factor influencing the exceedance rates of indices such as NH_3—N and BOD, and industries such as chemical raw materials, manufacturing, and pharmaceutical manufacturing are the main contributors to the exceedances of indices such as TP and NH_3—N, pollution prevention and control measures targeting industrial sources are relatively lacking in various regions. For example, the high proportion of pharmaceutical manufacturing in Yunnan Province, which increased in 2019 compared to 2018, may be an important factor contributing to the severe TP exceedance in this province. Similarly, the 5.68% increase in the TP exceedance rate in Shanghai in 2019 compared to 2018 may have been due to increased pharmaceutical manufacturing in the city without corresponding control measures. Therefore, we recommend that each region conducts assessments of industrial and key sectors based on their specific circumstances and implement corresponding control measures. This could effectively improve the water quality.

We also found that some areas had poor pollution control and even increased exceedance rates of certain indices. For example, in 2019, Chuzhou (Anhui Province) had high COD pollution, with a 27.78% increase in the COD exceedance rate compared to 2018. The pollution was concentrated in the Shuikou section. Kunming (Yunnan Province) had significant pollution in the Fumindaqiao section, the Xiguanqiao section in the Chuxiong Autonomous Prefecture also faced severe pollution, and the Huangdu section in Shanghai had a significant TP exceedance. For these sections (see Fig. S1 for the locations of the abnormal sections), we recommend establishing a list of point sources for investigation to determine any violations or potential pollution sources in the vicinity.

This study analyzed the drivers of the exceedances of six major water quality indices in 102 cities in the Yangtze River Basin at a more macroscopic scale and identified the risk factors for each city, thereby providing decision-making references for managers. However, this study has some limitations. First, as mentioned in the Materials and Methods section, owing to data quantity and scale

limitations, our study focused more on the impacts of spatial variations, and there was a relative lack of discussion on temporal factors. Future efforts will be dedicated to acquiring more comprehensive datasets to enhance the depth of our analysis. Second, the relationship between each driver and pollutant concentration was not explored in detail due to space limitations. Moving forward, we plan to employ machine learning techniques and other advanced methodologies to delve deeper into the interplay between various drivers and pollution concentrations.

4. Conclusions

This study reviewed the water quality of the Yangtze River Basin in 2018–2019 during the implementation of the WPPCAP and constructed a pollution inventory database for urban clusters in the Yangtze River Basin. Cities were classified into non-polluted, singly polluted, and compositely polluted. The exceedance drivers of six primary indices, including dissolved oxygen (DO), ammonia nitrogen (NH_3–N), chemical oxygen demand (COD), biochemical oxygen demand (BOD), total phosphorus (TP), and permanganate index (COD_{Mn}), were determined, and risk factor identification was performed for each city.

The analysis of the exceedance drivers revealed that agricultural factors were the main contributors to the exceedance of various indices in the basin from 2018 to 2019. Forest coverage also significantly affected the DO, NH_3–N, COD, and COD_{Mn} exceedance rates. Industrial factors, especially those related to pharmaceutical manufacturing and chemical raw materials and manufacturing, significantly influenced the NH_3–N, TP, and BOD exceedances. Animal husbandry was identified as an important factor causing NH_3–N, TP, and DO exceedances. Human activity-related factors significantly contributed to the COD, NH_3–N, and TP exceedances. In addition, climate factors strongly influenced the COD, BOD, and TP exceedances. Risk identification was conducted for 102 cities in the basin based on the top index exceedance drivers. Five risk factors were identified: human activities and industrial, agriculture and forest cover, climate, husbandry, and key polluting industry factors. The possible reasons for the exceedances caused by these drivers were analyzed, and targeted governance recommendations were proposed for urban agglomerations with high-risk ratings for each factor. Furthermore, cities with high pollution exceedances should use cities with effective pollution control measures as examples and develop tailored governance strategies based on local conditions.

By comprehensively analyzing the spatiotemporal drivers behind the pollution in the Yangtze River Basin from 2018 to 2019, the study analyzed the main possible causes of pollution and provided a risk assessment of the cities in the basin, which serves as a reference for the Yangtze River Basin governance planning until 2025, as mentioned in the “Key Basin Water Ecological Environment Protection Plan”. This study offers tailored governance suggestions for different types of cities, enabling managers to gain a clear and accurate understanding of policy directions.

In the future, we will further explore the detailed relationships between the spatiotemporal drivers of pollution in the Yangtze River Basin and the water quality indicators/pollutant concentrations in this study to make our conclusions more precise. Simultaneously, we intend to quantify the control measures proposed by the cities in the basin and further evaluate the impacts of the control measures on the water quality of the basin to provide greater assistance to basin managers.

CRediT authorship contribution statement

Yi-Lin Zhao: Conceptualization, Formal Analysis, Investigation, Data Curation, Methodology, Supervision, Validation, Writing - Original Draft, Writing - Review & Editing. **Shan-Shan Yang:** Conceptualization, Methodology, Supervision, Validation, Writing - Review & Editing, Funding Acquisition. **Mei-Yun Lu:** Methodology. **Han-Jun Sun:** Methodology. **Ji-Wei Pang:** Methodology. **Xiao-Dan Wang:** Investigation; **Da-Peng Zhou:** Investigation; **Ming Liang:** Investigation. **Nan-Qi Ren:** Conceptualization. **Jie Ding:** Conceptualization, Methodology, Supervision, Validation, Writing - Review & Editing, Funding Acquisition.

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.

Acknowledgements

The authors gratefully acknowledge financial support from the National Natural Science Foundation of China (Grant No. 52170073), the National Engineering Research Center for Bioenergy (Harbin Institute of Technology, Grant No. 2021A001), and the State Key Laboratory of Urban Water Resource and Environment (Harbin Institute of Technology) (Grant No. 2021TS03). We gratefully thank the financial support from the Joint Research program for ecological conservation and high-quality development of the Yellow River Basin (Grant No. 2022-YRUC-01-0204). We gratefully thank the contribution of the algorithm model and tool support by the artificial intelligence department of CECEP Digital Technology Co., Ltd. We gratefully acknowledge the support of the Heilongjiang Province Touyan Team.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2024.100412.

References

[1] N. Slimani, J.J. Jimenez, E. Guilbert, M. Boumaiz, J. Thioulouse, Surface water quality assessment in a semiarid Mediterranean region (Medjerda, Northern Tunisia) using partial triadic analysis, Environ. Sci. Pollut. Control Ser. 27 (2020) 30190–30198, https://doi.org/10.1007/s11356-020-09326-7.
[2] A. Kuriqi, A.N. Pinheiro, A. Sordo-Ward, M.D. Bejarano, L. Garrote, Ecological impacts of run -of -river hydropower plants ? Current status and future prospects on the brink of energy transition, Renewable Sustainable Energy Rev. 142 (2021), https://doi.org/10.1016/j.rser.2021.110833.
[3] A. Kuriqi, A.N. Pinheiro, A. Sordo-Ward, L. Garrote, Water-energy-ecosystem nexus: Balancing competing interests at a run-of-river hydropower plant coupling a hydrologic-ecohydraulic approach, Energy Convers. Manag. 223 (2020), https://doi.org/10.1016/j.enconman.2020.113267.
[4] X. Cheng, L. Chen, R. Sun, Y. Jing, An improved export coefficient model to estimate non-point source phosphorus pollution risks under complex precipitation and terrain conditions, Environ. Sci. Pollut. Control Ser. 25 (2018) 20946–20955, https://doi.org/10.1007/s11356-018-2191-z.
[5] C. Deng, L. Liu, D. Peng, H. Li, Z. Zhao, C. Lyu, et al., Net anthropogenic nitrogen and phosphorus inputs in the Yangtze River economic belt: spatiotemporal dynamics, attribution analysis, and diversity management, J. Hydrol. 597 (2021), https://doi.org/10.1016/j.jhydrol.2021.126221.
[6] Y. Han, Y. He, Z. Liang, G. Shi, X. Zhu, X. Qiu, Risk assessment and application of tea frost hazard in hangzhou city based on the random forest algorithm, Agriculture-Basel 13 (2023), https://doi.org/10.3390/agriculture13020327.
[7] Y. Ao, H. Li, L. Zhu, S. Ali, Z. Yang, The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling, J. Petrol. Sci. Eng. 174 (2019) 776–789, https://doi.org/10.1016/j.petrol.2018.11.067.
[8] B.R. Scanlon, S. Fakhreddine, R.C. Reedy, Q. Yang, J.G. Malito, Drivers of spatiotemporal variability in drinking water quality in the United States, Environmental Science & Technology 56 (2022) 12965–12974, https://doi.org/10.1021/acs.est.1c08697.
[9] S. Kumar, J. Pati, Assessment of groundwater arsenic contamination level in Jharkhand, India using machine learning, Journal of Computational Science 63 (2022), https://doi.org/10.1016/j.jocs.2022.101779.

[10] L. Li, J. Qiao, G. Yu, L. Wang, H.-Y. Li, C. Liao, et al., Interpretable tree-based ensemble model for predicting beach water quality, Water Res. (2022) 211, https://doi.org/10.1016/j.watres.2022.118078.
[11] A.N. Khiavi, M. Tavoosi, A. Kuriqi, Conjunct application of machine learning and game theory in groundwater quality mapping, Environ. Earth Sci. 82 (2023), https://doi.org/10.1007/s12665-023-11059-y.
[12] R.L. Olsen, R.W. Chappell, J.C. Loftis, Water quality sample collection, data treatment and results presentation for principal components analysis - literature review and Illinois River watershed case study, Water Res. 46 (2012) 3110–3122, https://doi.org/10.1016/j.watres.2012.03.028.
[13] C. Daou, M. Salloum, B. Legube, A. Kassouf, N. Ouaini, Characterization of spatial and temporal patterns in surface water quality: a case study of four major Lebanese rivers, Environ. Monit. Assess. 190 (2018), https://doi.org/10.1007/s10661-018-6843-8.
[14] W. Yang, Y. Zhao, D. Wang, H. Wu, A. Lin, L. He, Using principal components analysis and IDW interpolation to determine spatial and temporal changes of surface water quality of xin'anjiang river in huangshan, China, Int. J. Environ. Res. Publ. Health 17 (2020), https://doi.org/10.3390/ijerph17082942.
[15] D. Araya, J. Podgorski, M. Kumi, P.A. Mainoo, M. Berg, Fluoride contamination of groundwater resources in Ghana: country-wide hazard modeling and estimated population at risk, Water Res. 212 (2022), https://doi.org/10.1016/j.watres.2022.118083.
[16] X. Li, C. Wu, M.E. Meadows, Z. Zhang, X. Lin, Z. Zhang, et al., Factors underlying spatiotemporal variations in atmospheric PM2.5 concentrations in zhejiang province, China, Rem. Sens. 13 (2021), https://doi.org/10.3390/rs13153011.
[17] H. Guliyev, E. Mustafayev, Predicting the changes in the WTI crude oil price dynamics using machine learning models, Resour. Pol. 77 (2022), https://doi.org/10.1016/j.resourpol.2022.102664.
[18] N. Nordin, Z. Zainol, M.H.M. Noor, L.F. Chan, An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach, Asian Journal of Psychiatry 79 (2023), https://doi.org/10.1016/j.ajp.2022.103316.
[19] Z.-T. Zhao, H.-M. Cheng, S. Wang, H.-Y. Liu, Z.-M. Song, J.-H. Zhou, et al., SCC-UEFAS, an urban-ecological-feature based assessment system for sponge city construction, Environmental Science and Ecotechnology 12 (2022), https://doi.org/10.1016/j.ese.2022.100188.
[20] Y. Tong, W. Zhang, X. Wang, R.-M. Couture, T. Larssen, Y. Zhao, et al., Decline in Chinese lake phosphorus concentration accompanied by shift in sources since 2006, Nat. Geosci. 10 (2017) 507, https://doi.org/10.1038/ngeo2967.
[21] J. Xue, Q. Wang, M. Zhang, A review of non-point source water pollution modeling for the urban-rural transitional areas of China: research status and prospect, Sci. Total Environ. (2022) 826, https://doi.org/10.1016/j.scitotenv.2022.154146.
[22] H.X. Xu, X.X. Tan, J. Liang, Y.H. Cui, Q. Gao, Impact of agricultural non-point source pollution on river water quality: evidence from China, Frontiers in Ecology and Evolution 10 (2022), https://doi.org/10.3389/fevo.2022.858822.
[23] Y. Duan, H.Q. Jiang, X. Huang, W.H. Zhu, J. Zhang, B. Wang, et al., Evaluating nationwide non-point source pollution of crop farming and related environmental risk in China, Processes 11 (2023), https://doi.org/10.3390/pr11082377.
[24] M. Cui, Q.J. Guo, R.F. Wei, L.Y. Tian, Human-driven spatiotemporal distribution of phosphorus flux in the environment of a mega river basin, Sci. Total Environ. (2021) 752, https://doi.org/10.1016/j.scitotenv.2020.141781.
[25] C. Zhang, Z. Wen, J. Chen, An integrated model for technology forecasting to reduce pollutant emission in China's pulp industry, Resour. Conserv. Recycl. 54 (2009) 62–72, https://doi.org/10.1016/j.resconrec.2009.06.008.
[26] S. Cao, Y. Fei, X. Tian, X. Cui, X. Zhang, R. Yuan, et al., Determining the origin and fate of nitrate in the Nanyang Basin, Central China, using environmental isotopes and the Bayesian mixing model, Environ. Sci. Pollut. Control Ser. 28 (2021) 48343–48361, https://doi.org/10.1007/s11356-021-14083-2.
[27] H. Xu, S. Liu, Q. Xie, B. Hong, W. Zhou, Y. Zhang, et al., Seasonal variation of dissolved oxygen in Sanya Bay, Aquat. Ecosys. Health Manag. 19 (2016) 276–285, https://doi.org/10.1080/14634988.2016.1215743.
[28] J. Ding, Y. Jiang, L. Fu, Q. Liu, Q. Peng, M. Kang, Impacts of land use on surface water quality in a subtropical River Basin: a case study of the dongjiang River Basin, southeastern China, Water 7 (2015) 4427–4445, https://doi.org/10.3390/w7084427.
[29] X. Yi, Wastewater Treatment Plant Technology and Process Management, Chemical Industry Press, 2012, pp. 107–108.
[30] H.X. Xu, Q. Gao, B. Yuan, Analysis and identification of pollution sources of comprehensive river water quality: evidence from two river basins in China, Ecol. Indicat. 135 (2022), https://doi.org/10.1016/j.ecolind.2022.108561.
[31] L. Peng, X.Z. Deng, Z.H. Li, An extended input-output analysis of links between industrial production and water pollutant discharge in the Yangtze River Economic Belt, J. Clean. Prod. (2023) 390, https://doi.org/10.1016/j.jclepro.2023.136115.
[32] Z. Wu, W. Zhu, Y. Liu, P. Peng, X. Li, X. Zhou, et al., An integrated three-dimensional electrochemical system for efficient treatment of coking wastewater rich in ammonia nitrogen, Chemosphere 246 (2020), https://doi.org/10.1016/j.chemosphere.2019.125703.
[33] X. Zhang, H. Zhang, Z. Chen, D. Wei, Y. Song, Y. Ma, et al., Achieving biogas production and efficient pollutants removal from nitrogenous fertilizer wastewater using combined anaerobic digestion and autotrophic nitrogen removal process, Bioresour. Technol. (2021) 339, https://doi.org/10.1016/j.biortech.2021.125659.
[34] O. Vigiak, B. Grizzetti, A. Udias-Moinelo, M. Zanni, C. Dorati, F. Bouraoui, et al., Predicting biochemical oxygen demand in European freshwater bodies, Sci. Total Environ. 666 (2019) 1089–1105, https://doi.org/10.1016/j.scitotenv.2019.02.252.
[35] T. Gao, G. Gu, Q. Zhou, Water Pollution Control Engineering, fourth ed., Higher Education Press, 2014, pp. 8–9.
[36] G. Qiu, Y. Song, P. Zeng, S. Xiao, L. Duan, Phosphorus recovery from fosfomycin pharmaceutical wastewater by wet air oxidation and phosphate crystallization, Chemosphere 84 (2011) 241–246, https://doi.org/10.1016/j.chemosphere.2011.04.011.
[37] N. Chen, H. Hong, Nitrogen export by surface runoff from a small agricultural watershed in southeast China: seasonal pattern and primary mechanism, Biogeochemistry 106 (2011) 311–321, https://doi.org/10.1016/j.chemosphere.2011.04.011.
[38] H. Ye, X. Yuan, L. Han, J.B. Marip, J. Qin, Risk assessment of nitrogen and phosphorus loss in a hilly-plain watershed based on the different hydrological period: a case study in tiaoxi watershed, Sustainability 9 (2017), https://doi.org/10.3390/su9081493.
[39] D. Liu, L. Bai, X. Li, Y. Zhang, Q. Qiao, Z. Lu, et al., Spatial characteristics and driving forces of anthropogenic phosphorus emissions in the Yangtze River Economic Belt, China, Resour. Conserv. Recycl. (2022) 176, https://doi.org/10.1016/j.resconrec.2021.105937.
[40] D.J. Houlbrooke, S. Laurenson, Effect of sheep and cattle treading damage on soil microporosity and soil water holding capacity, Agric. Water Manag. 121 (2013) 81–84, https://doi.org/10.1016/j.agwat.2013.01.010.
[41] L. Xu, Y. Chen, Z. Wang, Y. Zhang, Y. He, A. Zhang, et al., Discovering dominant ammonia assimilation: implication for high-strength nitrogen removal in full scale biological treatment of landfill leachate, Chemosphere (2023) 312, https://doi.org/10.1016/j.chemosphere.2022.137256.
[42] S. Wu, M.M. Tang, Y. Wang, Z.W. Ma, Y.H. Ma, Analysis of the spatial distribution characteristics of livestock and poultry farming pollution and assessment of the environmental pollution load in Anhui province, Sustainability 14 (2022), https://doi.org/10.3390/su14074165.
[43] H. Bu, Y. Zhang, W. Meng, X. Song, Effects of land-use patterns on in-stream nitrogen in a highly-polluted river basin in Northeast China, Sci. Total Environ. 553 (2016) 232–242, https://doi.org/10.1016/j.scitotenv.2016.02.104.
[44] J. Xue, H. Li, J. Du, M. Zhou, Y. Mei, Temporal variation pollution source and decontamination characteristics of the Myriophyllum spicatum treatment pond, Ecol. Eng. 143 (2020), https://doi.org/10.1016/j.ecoleng.2019.105675.
[45] L. Huang, X. Han, X. Wang, Y. Zhang, J. Yang, A. Feng, et al., Coupling with high-resolution remote sensing data to evaluate urban non-point source pollution in Tongzhou, China, Sci. Total Environ. (2022) 831, https://doi.org/10.1016/j.scitotenv.2022.154632.
[46] S.M.B. Rahaman, M.S. Rahaman, A.K. Ghosh, D. Gain, S.K. Biswas, L. Sarder, et al., A spatial and seasonal pattern of water quality in the sundarbans river systems of Bangladesh, J. Coast Res. 31 (2015) 390–397, https://doi.org/10.2112/jcoastres-d-13-00115.1.
[47] C. Xu, X. Chen, L. Zhang, Predicting river dissolved oxygen time series based on stand-alone models and hybrid wavelet-based models, J. Environ. Manag. (2021) 295, https://doi.org/10.1016/j.jenvman.2021.113085.
[48] Q. Wang, Q. Chen, Simultaneous denitrification and denitrifying phosphorus removal in a full-scale anoxic-oxic process without internal recycle treating low strength wastewater, Journal of Environmental Sciences 39 (2016) 175–183, https://doi.org/10.1016/j.jes.2015.10.012.