

Article

Spatiotemporal Heterogeneity and Machine Learning Estimation of Key Water Quality Indicators in China's Coastal Waters Under Multi-Source Environmental Drivers

Huizhen Huang ¹, Zhangwen Su ^{1,*}, Yue Yang ¹, Xiaoxian Zheng ¹, Honghao Hu ², Kuangjun Zhao ³ and Canfeng Qiu ¹

¹ School of Petroleum and Chemical Engineering, Zhangzhou Institute of Technology, Zhangzhou 363000, China; huanghuizhen@t.fjzzit.edu.cn (H.H.); zzyyedu@126.com (Y.Y.); zhengxiaoxian@fjzzit.edu.cn (X.Z.); qcf0703@126.com (C.Q.)

² School of Equipment Manufacturing and Automobile Engineering, Zhangzhou Institute of Technology, Zhangzhou 363000, China; huhonghao@fjzzit.edu.cn

³ School of Artificial Intelligence, Zhangzhou Institute of Technology, Zhangzhou 363000, China; zhaokuangjun@fjzzit.edu.cn

* Correspondence: suzhangwen@fjzzit.edu.cn

Abstract

Water quality degradation in China's coastal waters poses a significant ecological challenge, yet the spatially varying predictors of key pollutants remain poorly quantified at the national scale. This study employed a Geographically Weighted Random Forest (GWRF) model integrating meteorological, marine, and socioeconomic data to analyze the associations between potential influencing factors and water quality variations in chemical oxygen demand (COD), dissolved inorganic nitrogen (DIN), and dissolved inorganic phosphorus (DIP) across China's coastal waters. Results reveal significant spatial heterogeneity and clustering patterns, with distinct "High-High" hotspots and "Low-Low" cold spots for each indicator. The GWRF model showed modest improvements over traditional global models (RF, XGBoost) under spatially structured cross-validation, with significant improvement only for COD. COD variations were jointly associated with meteorological and marine factors, whereas nutrients (DIN, DIP) showed stronger associations with marine environments. Spatialized SHAP effects indicated regional specificities: cropland proportion showed positive effects on COD in Bohai Bay but negative effects in eastern Shandong Peninsula; upper ocean heat content and coastal population density emerged as key contributing factors for DIN; sea surface salinity was closely linked to DIP distribution patterns. These findings provide a systematic quantification of factor contributions at the national scale, offering spatially explicit evidence for formulating differentiated pollution control strategies for China's nearshore waters.

Keywords: China's coastal waters; water quality; driving factors; Geographically Weighted Random Forest; spatial heterogeneity

Received: 10 June 2026

Revised: 8 July 2026

Accepted: 10 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As pivotal interfaces between land and sea, coastal zones host some of the most concentrated populations and economic activities globally. The ecological health of these regions is intrinsically linked to their sustainable development [1,2]. Nevertheless, persis-

tent discharges of pollutants from industrial, agricultural, and urban sources into coastal waters pose significant threats to marine environments [3–6]. In China, chemical oxygen demand (COD), dissolved inorganic nitrogen (DIN), and dissolved inorganic phosphorus (DIP) are critical parameters for evaluating coastal water quality. Elevated levels of these pollutants have led to the degradation of ecological functions in certain marine areas, prompting widespread concern among researchers and policymakers [1,7]. Consequently, a comprehensive understanding of pollution patterns and their underlying mechanisms in China's coastal waters are essential for formulating effective environmental management strategies.

In response to this need, the academic community has extensively explored the spatial distribution characteristics of water quality in China's coastal or nearshore areas. However, most of these studies have focused on specific local regions, such as individual estuaries (e.g., the Yangtze River Estuary and Pearl River Estuary), bays (e.g., Bohai Bay), or coastal waters of provinces [5,7–10]. These regional studies generally confirm that water quality pollution exhibits significant spatial heterogeneity and identify respective pollution hotspots. For instance, areas with high concentrations of industry and population have been repeatedly identified as zones of elevated pollution levels. Although these localized findings provide valuable case studies, they offer a fragmented and discontinuous perspective that fails to reveal the broader, continuous spatial patterns across the entirety of China's coastal zones. Consequently, there remains a lack of empirical research that systematically and holistically analyzes the spatial distribution of coastal water pollution on the national scale. This absence of a macroscopic perspective limits our ability to grasp the overall situation of environmental issues in China's coastal regions.

Traditional parametric models, such as multiple linear regression, offer interpretative advantages but fall short in handling the inherent high dimensionality and nonlinearity of environmental systems [7,11]. As a solution, machine learning models like Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), renowned for their robust predictive performance, have rapidly become mainstream analytical tools in environmental ecology [9,12,13]. However, these models often operate under a “global” assumption, positing that the relationships between driving factors and water quality indicators are homogeneous across the entire study area. This assumption overlooks the fundamental geographic principle of spatial non-stationarity, which states that the relationships between variables can vary with geographical location [7]. Consequently, global models may obscure critical regional characteristics, leading to systematic biases in interpreting driving mechanisms. To overcome the limitations of global models, hybrid models that integrate geographically weighted concepts with machine learning have emerged. By fitting a local model for each observation point in space, these hybrid models theoretically enable the capture of spatially heterogeneous effects of driving factors [7,11]. It is important to distinguish between traditional geographically weighted regression (GWR) and the machine-learning-based GWRF used in this study. GWR and its multiscale extension (MGWR) are parametric methods that assume local linearity; they estimate spatially varying coefficients but require the user to pre-specify functional forms and are sensitive to multicollinearity. In contrast, GWRF is a non-parametric approach that inherits the ability of random forests to capture complex non-linear relationships and high-order interactions among drivers without imposing linearity assumptions. Because the relationships between water quality and environmental drivers (e.g., temperature effects on algal growth and nutrient limitation thresholds) are likely non-linear and context-dependent, GWRF may be better suited for this study than GWR or MGWR. Furthermore, our primary goal is not to estimate interpretable linear coefficients but to maximize predictive accuracy as a foundation for SHAP-based importance analysis and spatial effect mapping. Although some studies have preliminarily validated the potential of such models in specific regions,

such as individual estuaries or bays [7,14], extending their application to the entirety of China's coastal waters and systematically comparing the driving mechanisms of various key pollutants remains an important research gap.

In view of this, we will develop a Geographically Weighted Random Forest (GWRF) model to systematically explore the spatial patterns of water quality pollution and their intricate driving mechanisms in China's coastal waters. Specifically, our research objectives are: (1) to reveal the spatial heterogeneity and clustering characteristics of COD, DIN, and DIP across China's coastal waters; (2) to validate the superiority of the GWRF model in handling spatial non-stationarity and compare its performance with traditional global machine learning models; and (3) to identify key driving factors and spatially visualize their effects using SHAP (SHapley Additive exPlanations) methods combined with geo-spatial techniques, thereby gaining deeper insights into the complex spatial variations in driving mechanisms. Through this study, we aim to understand the driving mechanisms and spatial disparities of water quality pollution in China's coastal waters more comprehensively, ultimately offering more targeted references for managing coastal water pollution and enhancing the efficiency and precision of environmental management strategies.

2. Materials and Methods

2.1. Study Area

China's coastal waters form part of the western Pacific marginal sea system, spanning from temperate to tropical climatic zones and encompassing the Bohai Sea, Yellow Sea, East China Sea, and the northern South China Sea (Figure 1). This extensive latitudinal gradient, combined with the East Asian monsoon, creates complex hydrodynamic conditions characterized by seasonal variations in northern regions [15] and typhoon dynamics in southern areas [16]. Between 2017 and 2024, the 11 coastal provinces, covering approximately 15% of China's land area, hosted about 45% of the population and contributed an average of 53% of national GDP annually (National Bureau of Statistics of China). Major rivers (Yangtze, Pearl) and economic hubs (Bohai Rim, Yangtze River Delta) have exacerbated organic pollution, hypoxia, and eutrophication in nearshore areas [7,10,15,17].

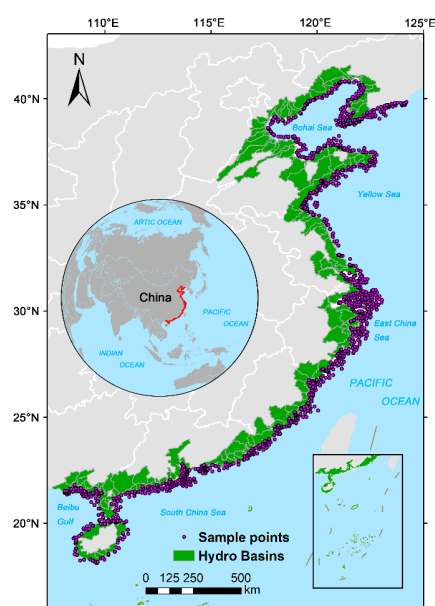


Figure 1. Study areas. The red line in the inset map represents the sampling points.

2.2. Data Collection and Processing

2.2.1. Coastal Water Quality

Water quality data were sourced from the National Marine Environmental Monitoring Center (NMEMC). To clarify the modeling framework, the specific spatio-temporal resolutions of the datasets are explicitly defined. Temporally, this study targets monthly average variations to capture seasonal dynamics. The dataset comprises 19,369 records collected monthly from May 2017 to August 2022 (primarily covering the months of May to December each year). Spatially, the modeling focuses on horizontal variability at discrete monitoring stations within the 12-nautical-mile coastal buffer zone across China's offshore coastal waters, including their latitude and longitude coordinates (Figure 1). The water quality observations represent surface water conditions, and the corresponding environmental variables were spatially matched to these specific point locations to capture kilometer-scale spatial heterogeneity along the coast. Three key water quality parameters were selected as response variables: Chemical Oxygen Demand (COD, mg/L), Dissolved Inorganic Nitrogen (DIN, mg/L), and Dissolved Inorganic Phosphorus (DIP, mg/L). These three parameters are core routine monitoring indicators in China's national marine environmental monitoring program and are the primary water quality metrics reported in annual government bulletins. Specifically, COD reflects the degree of organic pollution, while DIN and DIP are the main nutrients driving eutrophication in coastal waters. This dataset has been increasingly utilized in marine environmental studies [15,18]. The following preprocessing steps were applied: (1) records falling within a 12-nautical-mile (approximately 22.2 km) buffer zone along the coastline were retained. This boundary aligns with Chinese territorial sea definition under the United Nations Convention on the Law of the Sea (UNCLOS), providing a clear policy-relevant spatial scope (Figure 1); (2) for records marked as "below detection limit (<DL)", we replaced them with values calculated as $DL/(2\sqrt{2})$ following Helsel (2005) [19]. Specifically, the proportions of censored values in the final dataset were minimal: 0.38% for COD (54 records), 0% for DIN, and 3.65% for DIP (515 records). Given these low censoring rates (all < 5%), this simple substitution method is robust and introduces negligible bias to the dataset, as supported by environmental statistical theory [19]. According to the Chinese national standards for seawater quality monitoring (HJ 828-2017 for COD [20]; GB 17378.4-2007 [21] for DIN and DIP), the method detection limits are 4 mg/L for COD, 0.01 mg/L for DIN, and 0.003 mg/L for DIP.

2.2.2. Meteorological Conditions

Meteorological data were sourced from the ERA5 reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) via the Copernicus Climate Change Service (CDS) platform. Key meteorological indicators selected as potential explanatory variables included: 2-m temperature, total precipitation, mean sea level pressure, 10-m wind components (u and v), surface solar radiation downwards, and relative humidity. These variables have been identified as critical factors in previous studies [22–26]. Relative humidity data were obtained from pressure-level products, while the remaining six variables were derived from single-level products. Vector synthesis of the 10-m wind components was performed to derive wind speed (WS) and wind direction (WD). Using the "Extract Multi Values to Points" tool in ArcGIS 10.8 (Esri, Redlands, CA, USA), all meteorological variables were extracted to water quality sampling points based on geographic coordinates and corresponding monthly information, ensuring spatiotemporal alignment.

2.2.3. Marine Environment

Oceanographic data included sea surface temperature (SST), sea surface salinity (SSS), mixed layer depth (MLD), ocean heat content for the upper (OHCU), current speed,

and direction, sourced from the ORAS5 global ocean reanalysis [27]. Chlorophyll-a (CHL) concentration, representing phytoplankton biomass, was obtained from MODIS monthly composite products (NASA). Bathymetric data were sourced from the General Bathymetric Chart of the Oceans (GEBCO) with approximately 500 m resolution [28]. All oceanographic variables were extracted to water quality sampling points using the same ArcGIS 10.8 procedure.

2.2.4. Anthropogenic Activities

To investigate the impact of human activities on coastal water quality, several key indicators were selected. Population density data were sourced from the LandScan global population dataset (ORNL), and cropland coverage data from the European Space Agency's Climate Change Initiative (ESA CCI) land cover dataset. Using HydroBASINS level 7 watershed boundaries, the "Tabulate Intersection" tool in ArcGIS 10.8 was used to calculate population density (POP) and cropland area ratio (Rcrop) within each coastal terrestrial watershed. Additionally, the distance from each water quality sampling point to the nearest river mouth within its corresponding watershed (Dcoastal) was calculated to quantify pollutant transport pathways. All data sources, abbreviations, and spatio-temporal resolutions are summarized in Table 1, and the detailed summary statistics for each variable are presented in Table 2.

Table 1. The information of all variables.

Variable Types	Variable Name	Code	Unit	Resolution (Spatial/Tem- poral)	Description and Source
Response variable (Water quality in- dicators)	Chemical Oxygen Demand	COD	mg/L	-/monthly	National Marine Environmen- tal Monitoring Center (http://www.nmemc.org.cn/)(a ccessed on 1 October 2025)
	Dissolved Inor- ganic Nitrogen	DIN	mg/L		
	Dissolved Inor- ganic Phosphorus	DIP	mg/L		
Potential explana- tory variable (Air meteorologi- cal factors)	Relative air hu- midity	RH	%	0.25°/monthly	ERA5 monthly averaged data on pressure levels (<a href="https://cds.climate.coperni-
cus.eu/datasets/reanalysis-
era5-pressure-levels-monthly-
means?tab=overview">https://cds.climate.coperni- cus.eu/datasets/reanalysis- era5-pressure-levels-monthly- means?tab=overview) (ac- cessed on 1 October 2025)
	Total precipitation	TP	m		ECMWF Reanalysis 5th Gener- ation (ERA5) monthly aver- aged data on single levels (<a href="https://cds.climate.coperni-
cus.eu/datasets/reanalysis-
era5-single-levels-monthly-
means?tab=overview">https://cds.climate.coperni- cus.eu/datasets/reanalysis- era5-single-levels-monthly- means?tab=overview) (ac- cessed on 1 October 2025)
	2 m temperature	T2M	K		
	Mean sea level pressure	MSLP	hPa		
	10 m wind speed	WS	m/s		
	10 m wind direc- tion	WD	degree		
	Surface solar radi- ation downwards	SSRD	MJ/m ²		
Potential explana- tory variable (Marine environ- mental factors)	Mixed layer depth	MLD	m		Ocean Reanalysis System 5 (ORAS5) global ocean reanaly- sis monthly data (<a href="https://cds.climate.coperni-
cus.eu/datasets/reanalysis-">https://cds.climate.coperni- cus.eu/datasets/reanalysis-
	Sea surface salin- ity	SSS	PSU		
	Ocean heat con- tent for the upper	OHCU	GJ/m ²		

Potential explanatory variable (Human factors in coastal river basins)	Sea surface temperature	SST	°C		oras5?tab=overview) (accessed on 1 October 2025)
	Flow velocity of water	speed	m/s		
	The direction of water flow	direction	degree		
	Chlorophyll	CHL	mg/m ³		The Operational Mercator Ocean biogeochemical global ocean analysis and forecast system (https://data.marine.copernicus.eu/products) (accessed on 1 October 2025)
	Underwater digital elevation model	DEM	m	~500 m/static	General Bathymetric Chart of the Oceans (https://www.gebco.net)
	Population density along the coastal areas	POP	Pearson/km ²	~1 km/yearly	ORNL LandScan + HydroBASINS (level 07) (https://landscan.ornl.gov ; https://www.hydrosheds.org) (accessed on 1 October 2025)
Potential explanatory variable (Human factors in coastal river basins)	The ratio of cultivated land in the coastal river basins	Rcrop	%	~300 m/yearly	ESA CCI + HydroBASINS (level 07) (https://cds.climate.copernicus.eu/datasets/satellite-land-cover?tab=overview ; https://www.hydrosheds.org/products/hydrobasins) (accessed on 1 October 2025)
	The distance from the coastline	Dcoastal	m	-/static	HydroBASINA1:F22 (https://www.hydrosheds.org/products/hydrobasins) (accessed on 1 October 2025)

Table 2. Descriptive statistics of water quality and environmental variables.

Variables	Mean	Standard Deviation	Minimum	Maximum
COD	1.06	0.745	0.03	20.1
DIN	0.338	0.436	0.002	8.172
DIP	0.016	0.02	0.001	0.636
RH	74.804	10.936	0	91.74
TP	4.323	3.745	0	25.3
T2M	295.37	5.52	268.90	304.762
MSLP	1011.461	6.071	1000.995	1027.344
WS	2.68	1.638	0	11.928
WD	195.149	117.82	0	359.985
SSRD	17.604	3.641	7.086	25.293
MLD	13.828	4.149	0	54.774
SSS	29.099	5.259	0	34.455
OHCU	1.528	0.843	0.0386	8.003
SST	23.12	6.00	0	31.85

speed	0.092	0.102	0	0.876
direction	134.029	96.765	0	359.982
CHL	1.279	1.863	0	11.349
DEM	−11.755	14.829	−109	193.432
POP	690.184	568.333	1.463	3619.37
Rcrop	0.453	0.196	0	0.875
Dcoastal	5871.04	5246.612	0	19,552.3

2.3. Statistical Analysis

2.3.1. Spatial Autocorrelation Analysis

Before exploring driving mechanisms, we assessed whether water quality indicators exhibit spatial dependence. Global Moran's I was employed to evaluate overall spatial clustering [29]:

$$\text{Global Moran's } I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \times \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

where n is the number of water quality samples; x_i and x_j are the observed values of the water quality indicator at locations i and j ; $\bar{x}$ is the meaning of all observed values; w_{ij} represents the spatial weight between locations i and j in the spatial weights' matrix; the *Global Moran's I* value ranges is $[-1, 1]$.

Local spatial autocorrelation was analyzed using *Local Moran's I* to detect local association patterns [29]:

$$\text{Local Moran's } I_i = \frac{(x_i - \bar{x})}{S^2} \sum_{j \neq i} w_{ij} (x_j - \bar{x}) \quad (2)$$

where S^2 represents the variance of observed values. Results were categorized into four spatial association patterns: High–High (high-value observations surrounded by high-value neighbors, i.e., hotspots); Low–Low (low-value observations surrounded by low-value neighbors, i.e., coldspots); High–Low (high-value outliers in low-value areas); Low–High (low-value outliers in high-value areas). Analyses were conducted using the *spdep* package in R 4.5.2 (R Core Team, Vienna, Austria).

2.3.2. Mantel Test and Multicollinearity Assessment

The Mantel test was employed as an exploratory tool to assess the overall associations between each environmental variable and the three water quality indicator. This non-parametric method assesses correlation between two distance matrices [30]. Two distance matrices were constructed: one representing Euclidean distances of water quality indicators between sampling points, and another representing Euclidean distances of each candidate environmental driver. The Mantel correlation coefficient (r) was calculated, and permutation tests were used to determine significance. This analysis was performed using the *linkET* package in R 4.5.2. To account for the inflated Type I error rate due to multiple comparisons, all Mantel test p -values were adjusted using the Benjamini–Hochberg (BH) false discovery rate (FDR) procedure at a significance level of $q < 0.05$ [31]. Additionally, we assessed multicollinearity among the candidate predictors using the variance inflation factor (VIF) calculated from ordinary least squares regression (Section 2.3.3). Variables with $VIF > 10$ were considered collinear. A stepwise elimination procedure was therefore applied, iteratively removing the variable with the highest VIF until all remaining variables had $VIF \leq 10$.

2.3.3. Random Forest (RF)

Random Forest (RF) was introduced to analyze nonlinear impacts of environmental drivers on water quality. As an ensemble learning algorithm, RF constructs many decision trees and aggregates their predictions, effectively capturing complex variable interactions. Its “double randomness” mechanism (bootstrap sampling and random feature selection) provides resistance to overfitting [11,13]. RF was combined with SHAP analysis for interpretability. Model construction used the ranger and caret packages in R 4.5.2, with 10-fold cross-validation. Hyperparameters: number of trees = 500, minimum node size = 5, and mtry = one-third of predictors (rounded down).

2.3.4. eXtreme Gradient Boosting (XGBoost)

XGBoost was introduced as a second modeling approach to ensure robustness. Unlike RF which reduces variance through parallel independent trees, XGBoost minimizes bias through sequential iteration, with each new tree fitting residuals of the previous tree and incorporates L1/L2 regularization to prevent overfitting [10,13]. Models were implemented using the xgboost and caret packages with 10-fold cross-validation. Hyperparameters: learning rate = 0.1, maximum tree depth = 6, subsample ratio = 0.8, column sampling ratio = 0.8. Early stopping was applied after 50 iterations without improvement.

2.3.5. Geographically Weighted Random Forest (GWRF)

While global models (RF, XGBoost) can capture overall driving patterns, they assume homogeneous relationships across the study area. However, environmental influences on water quality often exhibit significant spatial non-stationarity [7,14]. To address this, we integrated spatial information directly into the RF framework. For any prediction point, the model assigns higher weights to neighboring training samples based on geographical proximity. A Gaussian kernel function was employed as the weighting function [11]:

$$w_{ij} = \exp[-(d_{ij}/b)^2] \quad (3)$$

where d_{ij} is the geographical distance between the i -th prediction point and the j -th training sample, and b is the bandwidth. The bandwidth determines the degree of locality: too small leads to overfitting, too large approximates global RF. In this study, we fixed the bandwidth at $k = 200$ nearest neighbors rather than using a fixed distance, to accommodate the uneven spatial distribution of sampling points. The value of $k = 200$ was chosen based on the average spatial density of sampling points to ensure sufficient local samples while effectively capturing spatial heterogeneity. For each prediction point, the k nearest neighboring samples were selected, and their spatial weights were calculated using the Gaussian kernel with b set as the maximum distance among the k nearest neighbors. These weighted samples were then used to train a local random forest model via the ranger package with the case.weights parameter, and the trained model was subsequently applied to predict the target point. GWRF used the same hyperparameter settings as global RF (ntree = 500, mtry = 6, nodesize = 5), and 10-fold spatial block cross-validation was employed for evaluation. GWRF was implemented in R 4.5.2 using the ranger, caret, and sp packages.

2.3.6. Model Interpretability Based on SHAP

To quantify the nonlinear contributions of each environmental driver and demystify the “black box” nature of machine learning models, the SHapley Additive exPlanations (SHAP) method was employed [32]. Based on Shapley values from cooperative game theory, SHAP calculates the average marginal contribution of a feature across all possible feature combinations, providing both global and local interpretability. In this study, SHAP values were calculated from the local GWRF model corresponding to each prediction point, with each sample’s SHAP values reflecting feature contributions specific to its

local spatial context. Global importance of each driver was ranked using the mean of absolute SHAP values. More importantly, local SHAP values for individual observations were mapped geographically to visualize spatially varying effects of key drivers, which is essential for understanding spatial non-stationarity in coastal water quality drivers. SHAP values are expressed in the same units as the response variable (mg/L for COD, DIN, and DIP).

2.3.7. Model Evaluation

Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). To avoid spatial leakage that may arise from random cross-validation in spatially autocorrelated data, we adopted a spatial block cross-validation approach. The study area was divided into 36 spatial grids of 20 km $\times$ 20 km, and all sampling points within each grid were assigned a unique grid identifier (NETID). These 36 grids were then evenly allocated into 10 folds, with each fold containing 3–4 grids. In each fold, the grids in that fold served as the test set, while the remaining nine folds served as the training set, resulting in a 10-fold spatial cross-validation. This strategy ensures that spatially adjacent samples do not appear in both the training and test sets simultaneously. The test set of each fold contained 1241 to 1720 samples, while the training set contained 12,382 to 12,861 samples. All models (RF, XGBoost, and GWRF) were evaluated using the same spatial block cross-validation scheme to ensure fair comparison. The technical roadmap of this study is illustrated in Figure 2.

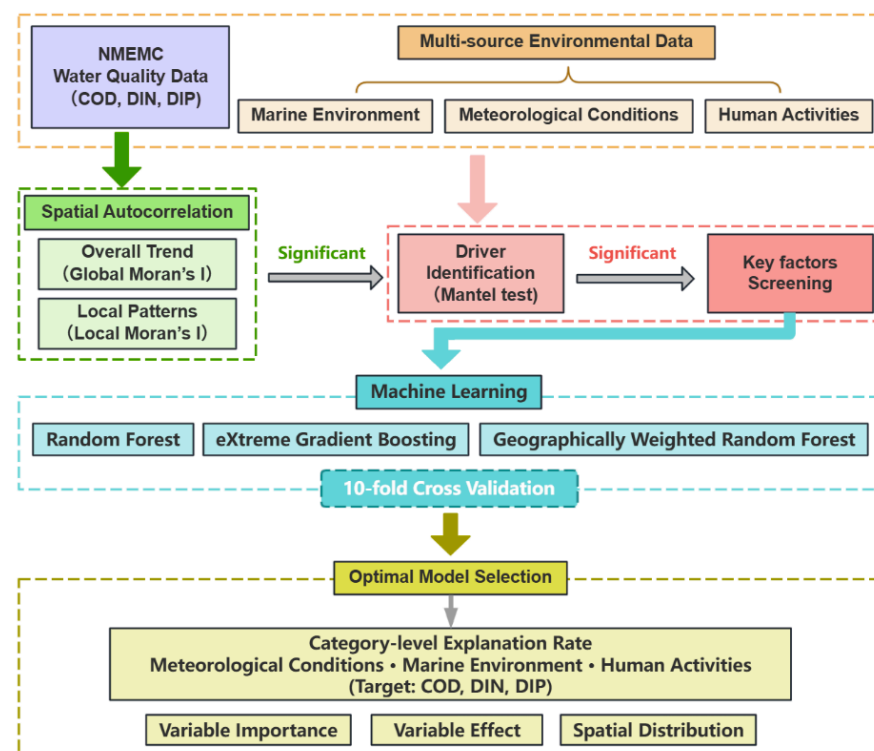


Figure 2. Technical roadmap of this research.

3. Results

3.1. Spatial Autocorrelation Characteristics of Water Quality Indicators

The results of Global Moran's I analysis indicated that during the period from May 2017 to August 2022, the three water quality indicators exhibited significant positive spatial autocorrelation at most points ($p < 0.05$), suggesting a clear clustering pattern in their spatial distribution (Figure 3). Among them, DIN showed the most significant and stable

spatial autocorrelation, with its Moran's I values being significant at all observed time points ($p < 0.05$), reaching a maximum of 0.93 in May 2021, indicating extremely strong spatial clustering. COD also demonstrated crucial spatial autocorrelation over most periods, except for two time points in September 2017 and November 2018, where it did not pass the significance test ($p > 0.05$). The spatial autocorrelation of DIP exhibited some variability, showing non-significant or even weak negative autocorrelation in certain months (such as September 2017, June 2018, and September 2019), but pivotal positive autocorrelation at most other time points.

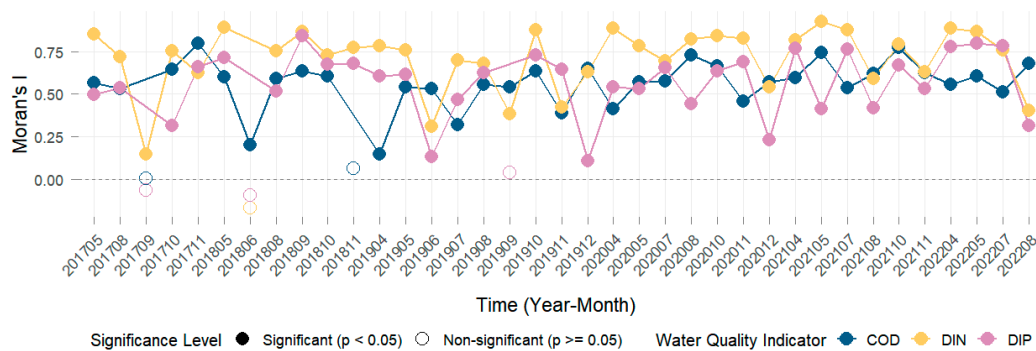


Figure 3. The line graph shows the distribution of Moran indices of spatial autocorrelation for COD, DIN, and DIP over time. Solid dots indicate significant results ($p < 0.05$); hollow dots indicate non-significant results. The dashed line indicates the zero reference line, representing no spatial autocorrelation.

The Local Moran's I analysis further elucidated the specific types of clustering and their spatial locations. The Local Moran scatterplot demonstrated that data points for the three indicators predominantly fell within the High–High and Low–Low quadrants, confirming that positive spatial autocorrelation was the predominant pattern (Figure 4). LISA cluster analysis vividly illustrates the distinct spatial clustering patterns for each of the three indicators, with notable variations in the clustering regions (Figure 5). Specifically, the High–High clusters for COD (comprising 28.7% of the data) were mainly concentrated along the Yangtze River Delta coastline and in the Bohai Bay area. In contrast, DIN exhibited High–High clusters (27.6%) primarily in the central coastal region. Similarly, DIP's High–High clusters (27.8%) were mostly found in the central coastal zone, though these did not extend to the southern part of the Bohai Bay coast. On the other hand, Low–Low clusters indicated areas with superior water quality. For COD, Low–Low clusters (52.6%) were widespread along the southern coastline. DIN's Low–Low clusters (62.2%) appeared both along the southern coast and at the fringes of the northern coast. Meanwhile, DIP's Low–Low clusters (59.3%) were predominantly observed along the Bohai Bay and the Jiaodong Peninsula coastline. Furthermore, the analysis identified Low–High clusters, which represented low-value areas surrounded by high-value regions, creating localized anomalies or island effects. These Low–High clusters were sporadically distributed across the three indicators, typically appearing at the peripheries of hotspots, indicating areas where water quality was significantly better than the surrounding polluted regions. These findings indicate that the spatial clustering patterns of different water quality parameters exhibit marked heterogeneity, reflecting the spatial variability in pollution sources and influencing factors.

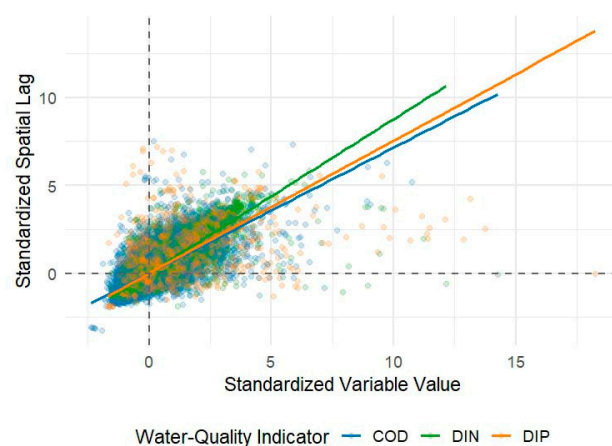


Figure 4. The global Moran scatter plot.

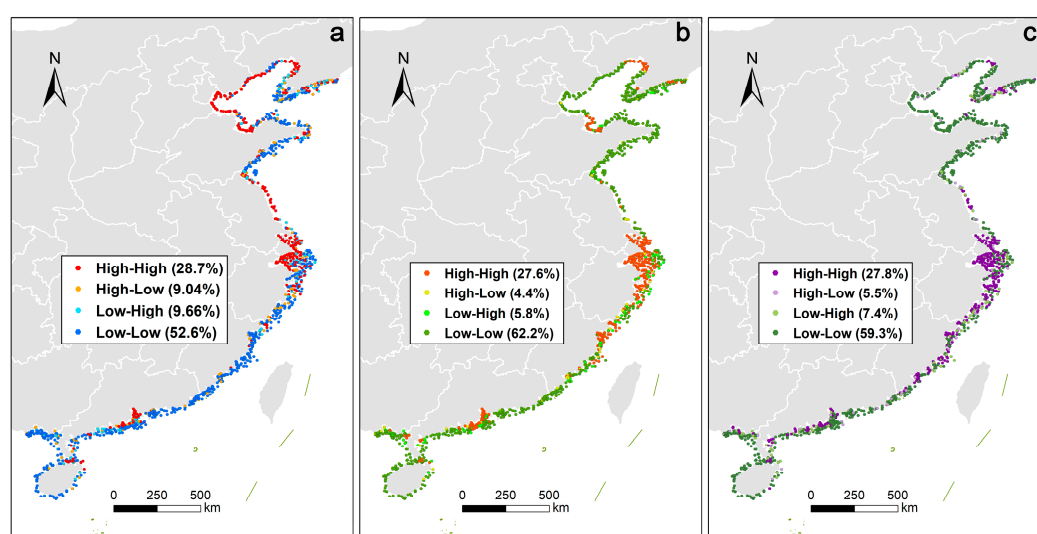


Figure 5. The LISA clustering diagram of COD (a), DIN (b), DIP (c).

3.2. Identification of Potential Predictors for Water Quality

The Mantel test revealed significant differences in the breadth and intensity of associations for the three water quality indicators (Figure 6). For COD, 9 out of 18 variables showed significant positive associations (FDR-adjusted $q < 0.01$): RH, T2M, MSLP, SSRD, SSS, SST, CHL, Rcrop, and Dcoastal. However, all correlation coefficients were below 0.2, indicating generally weak associations. For DIN, 7 variables exhibited significant positive associations ($q < 0.01$): TP, MLD, OHCU, SST, speed, direction, and POP, with OHCU showing the strongest correlation ($r = 0.177$). For DIP, 11 variables showed significant positive associations ($q < 0.01$): RH, TP, MSLP, WS, MLD, SSS, OHCU, SST, speed, Rcrop, and Dcoastal, with RH demonstrating the strongest correlation ($r = 0.260$). Notably, DIP was associated with the most extensive set of environmental factors, whereas COD exhibited the most limited set of significant associations. Detailed Mantel test results with FDR-adjusted q -values are presented in Figure 4. Furthermore, the maximum VIF among all 18 predictors was 6.37 (MSLP), well below the conventional threshold of 10 (Table 3), indicating that multicollinearity was not a major concern and all variables were retained for subsequent modeling.

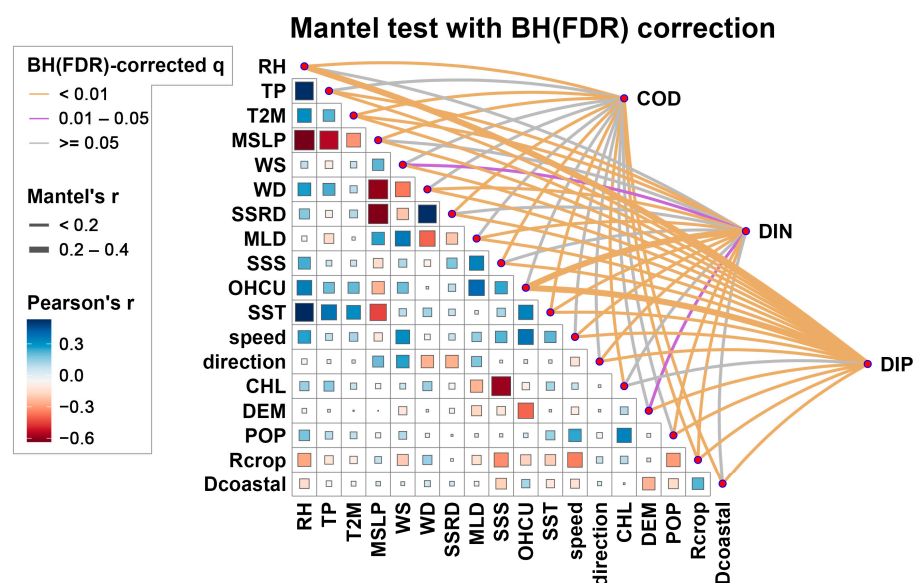


Figure 6. The result of mantel test.

Table 3. Variance inflation factor (VIF) values for all candidate predictors.

Variable	VIF
RH	2.67
TP	2.6
T2M	1.19
MSLP	6.37
WS	1.55
WD	2.03
SSRD	3.47
MLD	1.81
SSS	2.4
OHCU	2.05
SST	1.62
speed	1.53
direction	1.21
CHL	2.31
DEM	1.3
POP	1.38
Rcrop	1.44
Dcoastal	1.25

3.3. Performance Comparison of Machine Learning Models

The performance of the three machine learning models was evaluated using 10-fold spatial block cross-validation, with detailed fold-level results presented in Table 4. Overall, GWRF exhibited slightly better average RMSE and MAE than XGBoost and RF for most pollutants. However, the improvements were modest and not consistently observed across all folds. For COD, GWRF achieved a mean RMSE of 0.6416, compared to 0.6610 for XGBoost and 0.6372 for RF; for DIN, the mean RMSE values were 0.2850 (GWRF), 0.2892 (XGBoost), and 0.2873 (RF); for DIP, the values were 0.0166 (GWRF), 0.0166 (XGBoost), and 0.0163 (RF). To statistically assess these differences, paired t-tests were conducted on the fold-level RMSE between GWRF and the two global models (Table 5). GWRF showed significantly lower RMSE than XGBoost only for COD ($p = 0.018$), while no significant differences were detected for DIN ($p = 0.359$) or DIP ($p = 0.470$). Comparisons

between GWRF and RF revealed no significant differences for any pollutant ($p > 0.05$ for all). These results suggest that GWRF does not consistently outperform global models in predictive accuracy.

Table 4. Performance of RF, XGBoost, and GWRF models across 10 spatial cross-validation folds for COD, DIN, and DIP.

Response	Fold	RF			XGBoost			GWRF		
		R ²	RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE
COD	1	0.192	0.687	0.532	0.133	0.711	0.548	0.259	0.658	0.474
COD	2	0.122	1.208	0.653	0.080	1.237	0.669	0.134	1.200	0.622
COD	3	0.246	0.409	0.275	0.189	0.424	0.275	0.311	0.391	0.252
COD	4	−0.053	0.465	0.341	−0.131	0.482	0.360	−0.039	0.462	0.327
COD	5	0.196	0.764	0.377	0.198	0.763	0.374	0.178	0.772	0.382
COD	6	0.124	0.491	0.323	0.052	0.511	0.332	0.113	0.494	0.324
COD	7	−0.011	0.440	0.341	−0.002	0.438	0.333	0.100	0.415	0.302
COD	8	−0.006	0.544	0.368	−0.518	0.669	0.413	−0.445	0.652	0.380
COD	9	−1.297	0.846	0.634	−1.218	0.831	0.627	−1.399	0.864	0.638
COD	10	0.207	0.519	0.364	0.126	0.545	0.378	0.239	0.508	0.352
DIN	1	0.325	0.528	0.347	0.364	0.513	0.316	0.317	0.531	0.349
DIN	2	0.339	0.396	0.282	0.388	0.381	0.276	0.648	0.289	0.220
DIN	3	−0.347	0.179	0.116	−0.141	0.165	0.100	0.003	0.154	0.091
DIN	4	0.216	0.199	0.129	0.134	0.209	0.132	0.086	0.215	0.124
DIN	5	0.195	0.241	0.156	0.122	0.251	0.167	0.042	0.262	0.159
DIN	6	0.141	0.195	0.119	0.092	0.200	0.120	0.082	0.202	0.113
DIN	7	0.000	0.278	0.133	0.002	0.278	0.127	0.033	0.273	0.105
DIN	8	0.376	0.225	0.149	0.320	0.235	0.158	0.415	0.218	0.144
DIN	9	0.528	0.384	0.266	0.468	0.407	0.290	0.475	0.404	0.282
DIN	10	0.390	0.247	0.143	0.365	0.252	0.143	0.099	0.301	0.163
DIP	1	0.043	0.023	0.015	0.127	0.022	0.013	0.138	0.022	0.012
DIP	2	−0.146	0.019	0.015	−0.098	0.019	0.014	0.386	0.014	0.011
DIP	3	0.105	0.009	0.006	0.060	0.009	0.006	−0.072	0.010	0.006
DIP	4	0.151	0.023	0.010	0.126	0.024	0.010	0.161	0.023	0.010
DIP	5	0.150	0.013	0.009	0.062	0.013	0.009	−0.746	0.018	0.011
DIP	6	0.098	0.022	0.008	0.098	0.022	0.008	0.116	0.022	0.008
DIP	7	0.059	0.011	0.007	0.073	0.011	0.007	0.043	0.011	0.006
DIP	8	0.295	0.019	0.012	0.251	0.019	0.012	0.313	0.018	0.011
DIP	9	0.518	0.012	0.010	0.365	0.014	0.010	0.184	0.016	0.011
DIP	10	−0.110	0.012	0.008	−0.299	0.013	0.009	−0.009	0.012	0.008

Table 5. Paired *t*-test results comparing GWRF with XGBoost and RF across 10 spatial cross-validation folds based on RMSE.

Comparison	Response	Mean_Diff	Sd_Diff	t_Stat	p_Value
GWRF vs. XGBoost	COD	−0.0194	0.0249	−2.4541	0.0183
GWRF vs. XGBoost	DIN	−0.0042	0.0358	−0.3725	0.3591
GWRF vs. XGBoost	DIP	−0.0001	0.0024	−0.0778	0.4699
GWRF vs. RF	COD	0.004	0.039	0.352	0.634
GWRF vs. RF	DIN	−0.002	0.042	−0.168	0.435
GWRF vs. RF	DIP	0.0003	0.0028	0.2799	0.6071

3.4. Interpretability Analysis of GWRF

After confirming the superior performance of the GWRF model, we conducted an interpretability analysis to uncover the contributions of different categories of driving factors to water quality variations. We analyzed the specific effects and importance of individual drivers on water quality metrics. SHAP effect analysis based on 10-fold cross-validation (Figure 7) and importance ranking (Figure 8) revealed distinct key drivers for different pollutants. The SHAP analysis revealed distinct importance patterns across the three pollutants. For COD, CHL (Mean_Abs = 0.0953) and Rcrop (0.0856) were the two dominant predictors, substantially exceeding all other variables, followed by SSS (0.0812). Notably, all three variables exhibited negative mean SHAP values, suggesting that higher values of these predictors were generally associated with lower COD predictions, though their positive maxima (CHL: 2.15, Rcrop: 1.58, SSS: 2.30) indicate that positive effects emerged in specific localized regions. For DIN, Rcrop (0.0503) and SSS (0.0501) ranked highest, followed by POP (0.0383) and OHCU (0.0373). While Rcrop and POP showed negative mean effects (−0.0372 and −0.0178, respectively), OHCU exhibited a small but consistent positive mean effect (0.0079), which may be associated with thermal stratification in certain areas. For DIP, SSS (0.00237) was the most influential predictor, followed by OHCU (0.00183) and Rcrop (0.00154). Overall, the magnitude of SHAP effects for DIP was notably smaller than those for COD and DIN (by approximately one order of magnitude), suggesting that DIP variations are more weakly associated with individual factors and may reflect complex interactions among multiple variables.

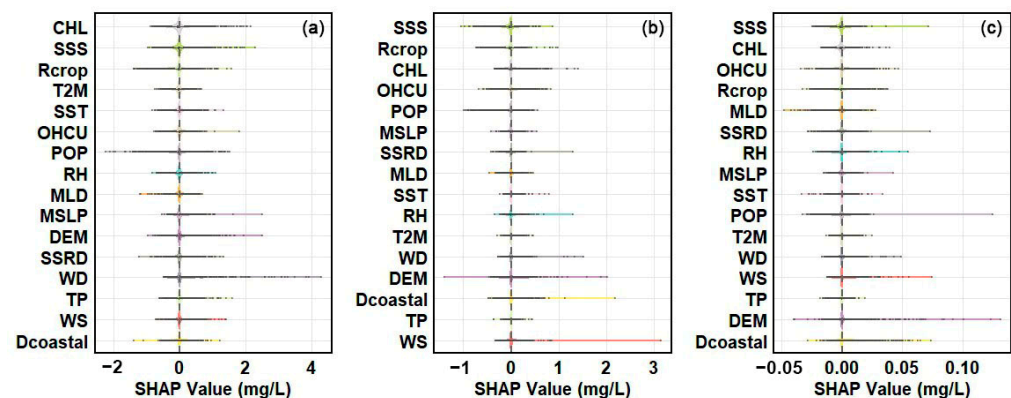


Figure 7. GWRF-based SHAP Bee Swarm Plot of Driving Factors for COD (a), DIN (b), and DIP (c).

To further investigate the spatial heterogeneity of these key driving factors and their potential applications in pollution regulation, we evaluated the cumulative contributions of different variable combinations to model explanatory power (Figure 8). The results indicated that a small number of core variables could account for most of the model's predictive capacity. For COD, the first four variables (CHL, Rcrop, SSS, and POP) collectively achieved a cumulative contribution of 86.9%, while the addition of DEM as the fifth variable pushed the cumulative R^2 to 93.7%, exceeding the 90% threshold. For DIN, the first four variables (Rcrop, SSS, POP, and OHCU) together explained 92.4% of the variance, with agricultural activities and marine dynamic processes emerging as key contributing factors. In the case of DIP, although individual variable contributions were relatively dispersed, the first five variables (SSS, OHCU, CHL, Rcrop, and SSRD) accounted for 87.1%, and the inclusion of Dcoastal as the sixth variable reached 96.2%, surpassing the 90% threshold. Based on these findings, we identified core variable combinations that cumulatively explained over 90% of the variance and performed spatial mapping of key driving factor effects (Figure 9) to visually identify hotspots of pollution and their spatial correlations.

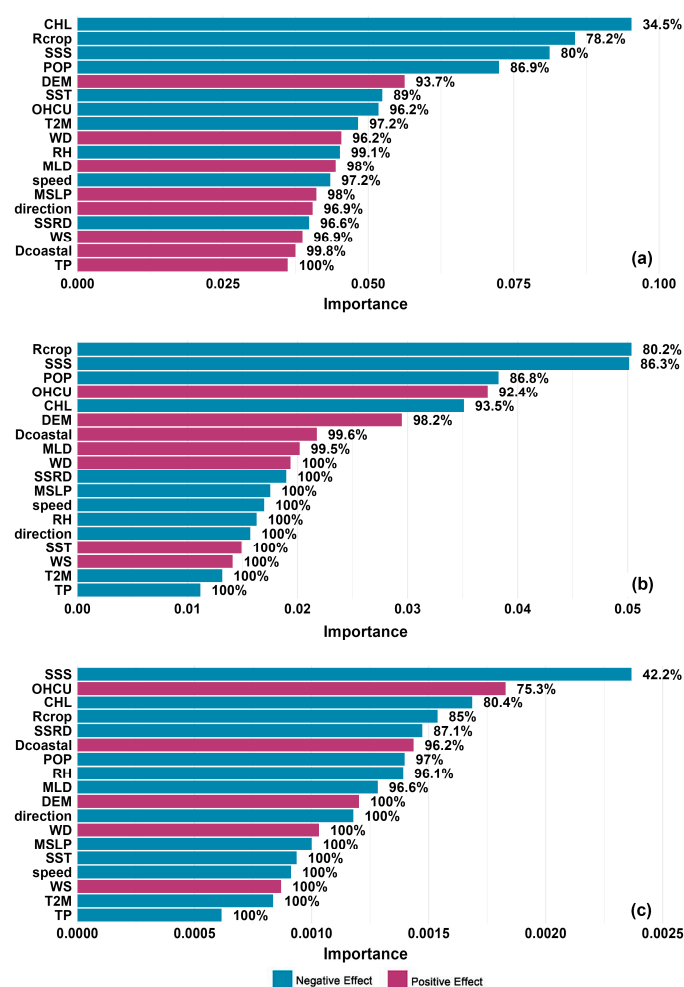


Figure 8. Variable importance ranking for (a) COD, (b) DIN and (c) DIP obtained from SHAP analysis of the GWRF.

The spatial SHAP effects of key factors for COD, DIN, and DIP exhibited significant indicator-specific differences in effect direction, spatial aggregation, and magnitude. For COD, CHL, Rcrop, SSS, and POP were predominantly negative, with negative proportions from 69% to 78% and mean SHAP values between -0.024 and -0.062 . Positive effects were primarily concentrated in the Bohai Bay and the Yangtze River Estuary–Hangzhou Bay area. CHL showed positive peaks up to 1.19 in the Liaodong Bay and the Yangtze River Estuary, while negative effects as low as -0.63 prevailed in the central Yellow Sea and the East China Sea shelf. Rcrop exhibited positive effects up to 0.54 in the Bohai Bay and the Shandong Peninsula coast, but negative effects down to -0.88 across most other regions. SSS displayed positive effects up to 0.75 along the Shandong Peninsula coast, contrasting with negative effects as low as -0.84 around the Yangtze River Estuary. POP showed positive effects up to 0.78 in the Bohai Bay and the Yangtze River Delta, with negative effects down to -0.85 in open shelf waters. DEM was the only variable exhibiting predominantly positive effects (73%), with positive values up to 1.07 in shallow waters and negative values only appearing in deeper waters of the outer East China Sea shelf and the northern South China Sea (as low as -0.37).

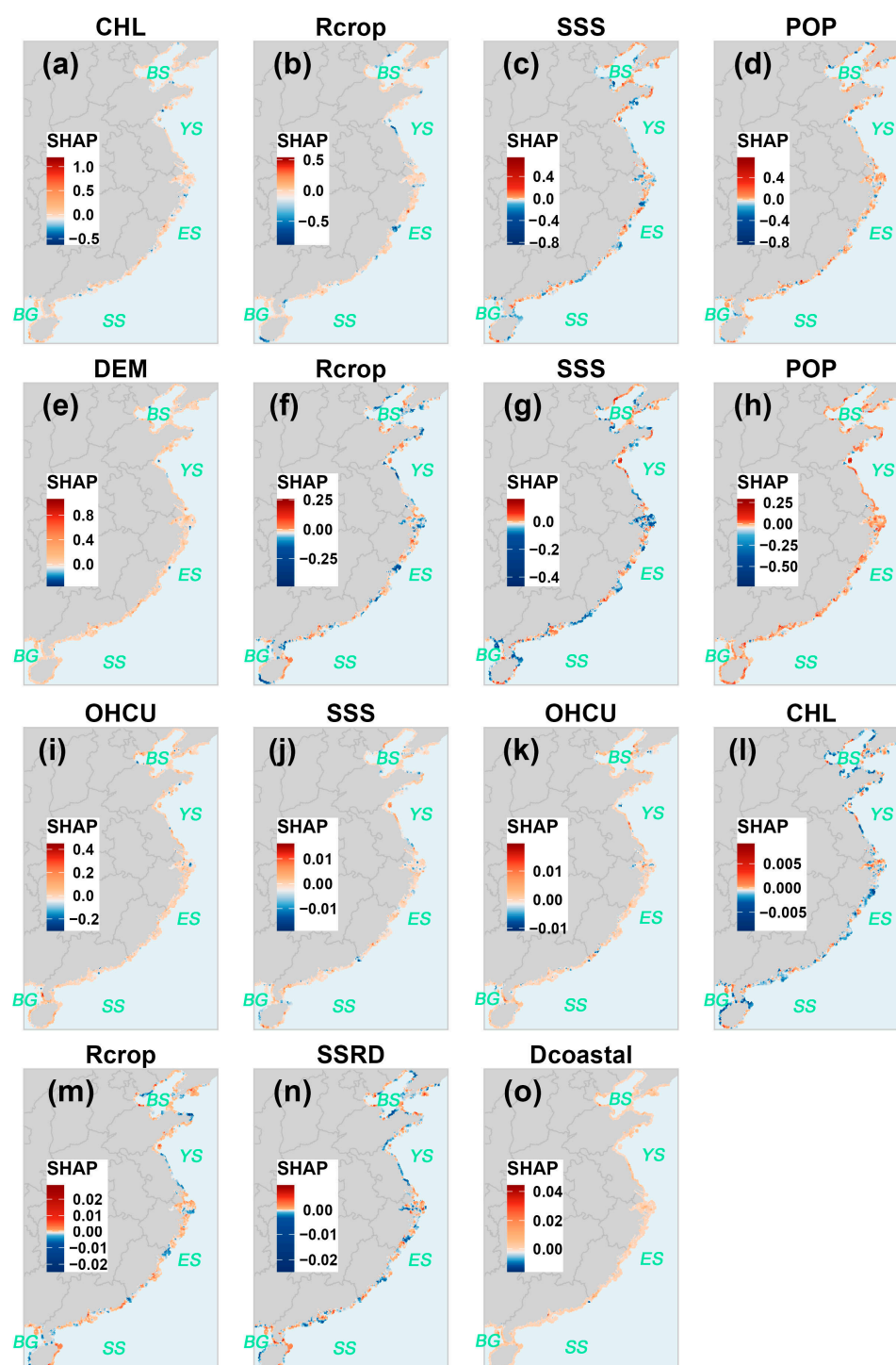


Figure 9. Spatial SHAP effects of key drivers for COD, DIN, and DIP across China's coastal waters. Panels (a–o) show the spatial distribution of SHAP values for selected predictors, with warm colors (orange to dark red) indicating positive effects (increasing the predicted concentration) and cool colors (light to dark blue) indicating negative effects (decreasing the predicted concentration). Specifically, panels (a–e) present the effects for COD of CHL (a), Rcrop (b), SSS (c), POP (d), and DEM (e); panels (f–i) present the effects for DIN of Rcrop (f), SSS (g), POP (h), and OHCU (i); and panels (j–o) present the effects for DIP of SSS (j), OHCU (k), CHL (l), Rcrop (m), SSRD (n), and Dcoastal (o). The spatial heterogeneity of these effects highlights the region-specific driving mechanisms of different water quality indicators, providing spatially explicit evidence for differentiated pollution control strategies. Abbreviations used in the map: BS is Bohai Sea; YS is Yellow Sea; ES is East China Sea; SS is South China Sea; BG is Beibu Gulf.

For DIN, Rcrop, SSS, and POP were also predominantly negative, with negative proportions from 72% to 86% and mean SHAP values between -0.019 and -0.039 , while OHCU showed a slight predominance of positive effects (58%, mean = 0.005). Rcrop exhibited positive effects up to 0.26 in the Bohai Bay and the Shandong Peninsula coast, but negative effects down to -0.48 across most other regions. SSS showed positive effects up to 0.16 in the Bohai Bay and the Shandong Peninsula coast, with negative effects as low as -0.47 in the Yangtze River Estuary. POP displayed positive effects up to 0.29 in the Bohai Bay, the Yangtze River Estuary, and the Pearl River Estuary, but negative effects down to -0.73 in open shelf waters. In contrast, OHCU exhibited a distinct spatial pattern: positive effects up to 0.45 were distributed in the Bohai Sea, Yellow Sea, and northern East China Sea, while negative effects appeared in the outer East China Sea shelf and the northern South China Sea (as low as -0.30).

For DIP, all six factors exhibited extremely low effect magnitudes, with mean SHAP values from -0.001 to 0.001 and standard deviations below 0.003 , approximately one to two orders of magnitude lower than those for COD and DIN. SSS, CHL, and Rcrop were predominantly negative, with negative proportions from 73% to 79% and means between -0.00102 and -0.00071 . SSS showed positive effects up to 0.016 in the Bohai Bay and the Shandong Peninsula coast, and negative effects as low as -0.019 in the Yangtze River Estuary. CHL exhibited positive effects up to 0.009 in the Bohai Bay and patchy areas of the northern South China Sea, with negative effects down to -0.009 across most other regions. Rcrop displayed positive effects up to 0.029 in the Bohai Bay and the Shandong Peninsula coast, and negative effects as low as -0.025 in the Yangtze River Estuary and Zhejiang coastal waters. OHCU and SSRD showed roughly balanced positive and negative effects with negligible means (0.00014 and -0.00014 , respectively), and their spatial patterns were scattered without clear regional aggregation. In stark contrast, Dcoastal was the only factor showing a clear predominance of positive effects (68%), with positive values up to 0.044 concentrated in the Yangtze River Estuary, Fujian coastal waters, and the Pearl River Estuary, while negative effects only appeared in offshore open waters (as low as -0.016).

4. Discussion

4.1. Spatial Heterogeneity and Aggregation Patterns of Coastal Water Quality Pollution

The findings of this study clearly demonstrate that water quality pollution in China's coastal areas is not randomly distributed but exhibits significant spatial heterogeneity and clustering patterns, reflecting the highly non-stationary nature of pollution sources and marine environmental processes. From a temporal perspective, COD, DIN, and DIP consistently exhibited strong positive spatial autocorrelation during most periods (Figure 3), confirming that clustering effect is a prevalent characteristic of coastal water quality in China [13,33–35]. The levels of spatial autocorrelation for these three water quality indicators showed notable fluctuations, indicating the substantial influence of seasonal variations and episodic events on China's coastal water systems [33,36–38].

Local spatial autocorrelation analysis further delineated the internal variability of this clustering effect, with distinct hotspots and coldspots for different pollutants, suggesting the complexity of underlying driving mechanisms. COD's High–High clusters (hotspots) were highly concentrated in the Yangtze River Delta and Bohai Bay regions. This pattern aligns closely with the core areas of two major economic zones in China, where industrial activities are dense, populations are large, and port operations are frequent, contributing to substantial discharges of industrial and domestic wastewater. These pollutants accumulate in estuaries and adjacent waters via major river systems such as the Yangtze, Liao, and Hai Rivers, forming pronounced pollution hotspots [39–41]. The hotspots for DIN and DIP not only encompassed the Yangtze River Estuary but also extended towards central coastal areas, indicating potential differences in dominant pollu-

tion sources. Compared to COD, nutrients (DIN/DIP) have more diverse origins, including agricultural runoff and aquaculture effluents, in addition to industrial and domestic point-source emissions [42,43]. Central coastal regions in China serve as significant agricultural and aquaculture hubs, which may be associated with their status as nutrient hotspots [44,45]. Furthermore, the identification of Low–Low clusters (cold spots) provides valuable insights into marine environmental protection. For instance, COD’s cold spots were widely distributed along the southern coast, while DIN and DIP cold spots appeared simultaneously in the southern and northern coastal peripheries. These areas exhibited relatively better water quality, possibly due to lower human activity intensity [41,46,47] or more favorable hydrodynamic conditions (such as stronger current dilution effects) [41,48]. The presence of these ecological oases highlights the association between human activities in influencing coastal water quality.

4.2. The Necessity and Efficacy of GWRF in Modeling Spatially Non-Stationary Relationships

We used RF and XGBoost as baseline models to evaluate the performance of GWRF in our study. Given the significant spatial heterogeneity observed in the study area, which contradicts the constant assumption of global models, the introduction of GWRF aimed to quantify the improvement in model performance by accounting for spatial non-stationarity. The GWRF model exhibited modest improvements in fitting the relationships between the three key water quality indicators (COD, DIN, and DIP) and their drivers relative to the two global baseline models in terms of prediction accuracy. However, these improvements were not consistently observed across all pollutants or all cross-validation folds. As reported in Section 3.3, GWRF showed significantly lower RMSE than XGBoost only for COD ($p = 0.018$), with no significant differences detected for DIN or DIP, and no significant differences compared to RF for any pollutant. Despite the modest predictive gains, the primary value of GWRF lies in its ability to provide spatially localized interpretation through SHAP effects, which is not attainable with global models. This aligns with previous studies that have highlighted the value of geographically weighted approaches in elucidating spatial heterogeneity in environmental systems [7,14,49].

4.3. Interpreting the Spatial Associations of Water Quality Factors

Our SHAP-based spatial analysis revealed distinct association patterns between environmental factors and three water quality indicators, with key predictors varying by pollutant. COD was associated with both terrestrial inputs and biogeochemical processes, whereas DIN and DIP were more strongly associated with marine physical and biogeochemical conditions—consistent with known pollution dynamics in China’s coastal waters. Studies in the central Bohai Sea found nutrients more strongly associated with internal processes, such as phytoplankton uptake and sediment–water interface exchanges, rather than simple terrestrial accumulation [50], consistent with the weaker and more spatially dispersed SHAP patterns we observed for these nutrients. In contrast, COD’s SHAP associations reflected a more balanced combination of land-based and marine-related factors. Specifically, wind speed and precipitation affect terrestrial inputs and mixing processes, which together with ocean currents and estuarine transport shape COD’s spatio-temporal distribution [47,48]. Seasonal COD patterns in the Bohai Sea support this: summer maxima are linked to terrestrial sources, while winter gradients are attributed to ocean circulation [50]. Our spatial SHAP results extend these findings by pinpointing where these associations are spatially strongest.

Differences in association patterns among COD, DIN, and DIP are consistent with their distinct roles in coastal water quality. COD, as a comprehensive organic pollution indicator, derives from both land-based industrial/domestic wastewater and autochthonous organic matter from phytoplankton production—a dual origin explaining its spatial

associations with terrestrial inputs (cropland, population density) and a biological process indicator (chlorophyll-*a*). DIN, directly assimilated by phytoplankton, shows strong associations with internal cycling (biological uptake and nitrification–denitrification), accounting for its generally weaker SHAP effect magnitudes than Chemical oxygen demand and for upper ocean heat content (a physical stratification indicator) emerging as a more prominent correlate. DIP shows extremely weak effect magnitudes, which may reflect its tendency to bind with suspended particles and settle nearshore; its spatial distribution is governed more by abiotic processes (physical mixing, adsorption–desorption, sediment release) than by biological uptake, making distance to coast the strongest correlate among all factors examined.

The spatially explicit SHAP patterns further resolve regional heterogeneities beyond global models. For chemical oxygen demand, the positive cropland proportion association in Bohai Bay and the Yangtze River Estuary–Hangzhou Bay aligns with evidence that agricultural watersheds are major organic matter sources to nearshore waters [50,51]. However, the negative association in the eastern Shandong Peninsula and the Pearl River Estuary indicates more complex regional specificity. In the eastern Shandong Peninsula, intricate biogeochemical interactions may alter or mask the agricultural signal [52]; in the Pearl River Estuary, pollutants may originate more from regional transport than local croplands, creating a source–sink mismatch that yields a negative SHAP association [53]. Chlorophyll-*a* showed patchy positive associations in Bohai Bay, the Yangtze River Estuary, and southern Hainan Island—consistent with prior observations of high chlorophyll-*a* and primary productivity in Bohai Bay [54] and likely reflecting enhanced nutrient inputs that promote phytoplankton growth [55].

For DIN, SHAP patterns indicate combined influences of agricultural sources, marine physical processes, and anthropogenic emissions. The cropland proportion–DIN association resembles that of COD but with markedly reduced magnitude, consistent with agricultural fertilizer runoff as a nitrogen source whose spatial signature is often reshaped by marine processes [56]. Upper ocean heat content showed positive associations in the Bohai Sea and Hainan Island, consistent with the hypothesis that thermal stratification may inhibit vertical mixing and retains nutrients in surface waters [57,58]. Population density exhibited positive associations primarily in coastal urban zones, which may reflect spatial coincidence of anthropogenic nitrogen emissions and elevated DIN—also observed in Lai-zhou Bay, where riverine anthropogenic emissions interact with natural factors to shape nutrient distributions [41].

For DIP, SHAP associations indicate stronger ties to physical and geochemical factors than to biological ones. Sea surface salinity–DIP associations shared overlapping positive zones with COD but also distinct negative zones, reflecting phosphorus cycling complexity. In estuarine areas, negative salinity–DIP associations typically reflect spatial coupling of freshwater discharge and phosphorus inputs; positive associations in coastal waters may relate to biological consumption or particle adsorption and sedimentation [59]. Distance to coast showed the strongest and most spatially coherent positive association with DIP, highlighting land–ocean distance as a key distribution factor. Surface downward solar radiation exhibited a weak positive association, consistent with photochemical processes that may enhance phosphorus release from sediments [59,60].

By spatializing SHAP effects, our analysis moves beyond global model interpretations to reveal the local spatial structure of factor–response associations. These refined spatial patterns provide direct evidence for developing targeted management strategies. In hotspot areas where cropland proportion shows positive associations, such as the Bohai Bay and the Yangtze River Estuary, agricultural nonpoint source pollution control should be prioritized. In coastal urban belts where population density and upper ocean heat content show notable positive associations, enhanced regulation of point-source emissions

and attention to stratification-related nutrient retention may be warranted. In areas where chlorophyll-a associations are patchy, management attention should be directed toward internal nutrient cycling and eutrophication mitigation.

4.4. Limitations and Future Research Directions

Despite the valuable insights gained from this study, there are limitations in both methodology and consideration of influencing factors, which point to important directions for future research. On the methodological front, our study primarily uncovered statistical correlations between driving factors and water quality indicators but fell short of fully elucidating the underlying causal mechanisms. For example, while the model identified a relationship between upper ocean heat content and DIN concentrations, it could not confirm thermal stratification as a direct physical pathway. Future research should explore integrating machine learning with process-based models or structural equation modeling to more comprehensively dissect these complex causal chains. In terms of influencing factors, there is room for expanding the scope of our driver set. First, regarding temporal resolution, the current data predominantly consists of monthly or annual averages, which are insufficient for capturing the pulse effects of short-term extreme events such as typhoons and heavy rainfall on water quality. Future studies should incorporate higher-resolution time series data and employ spatiotemporal models to better understand these dynamic processes. Second, concerning the types of driving factors, our study mainly focused on meteorological, marine, and anthropogenic pressures but did not fully account for indirect drivers such as socioeconomic policies. For instance, how tightening environmental regulations or adjustments in agricultural subsidy policies influence human behavior and ultimately affect water quality remains an area ripe for exploration. Incorporating these macro-level factors into models would provide a more holistic understanding of the deeper causes behind water quality changes.

5. Conclusions

This study aims to systematically reveal the spatial patterns and complex driving mechanisms of water quality in China's coastal waters. To achieve this, we developed a Random Forest model integrated with geographically weighted concepts to dissect the spatial non-stationarity of driving factors. The main conclusions are as follows: Firstly, water quality pollution in China's coastal waters exhibits significant spatial heterogeneity and clustering effects. COD pollution hotspots are concentrated in industrial and densely populated regions such as the Yangtze River Delta and Bohai Bay, while nutrient (DIN, DIP) hotspots extend towards central coastal areas where agriculture and aquaculture are intensive. This spatial differentiation provides precise targets for region-specific management strategies. Secondly, the GWRF model demonstrated modest improvements in predictive performance over global models (RF and XGBoost) under spatially structured cross-validation. Paired *t*-tests indicated that GWRF showed significantly lower prediction errors than XGBoost only for COD ($p = 0.018$), while no significant differences were detected for DIN or DIP. These results suggest that the primary value of GWRF lies not in maximizing predictive accuracy, but in its capacity to reveal spatially varying factor-response associations through localized SHAP effects—a capability that global models inherently lack. Third, the spatial SHAP analysis identified distinct association patterns across the three pollutants. COD showed predominantly negative associations with chlorophyll-a, cropland proportion, sea surface salinity, and population density, with positive effects concentrated in the Bohai Bay and the Yangtze River Estuary–Hangzhou Bay area; water depth was the only factor exhibiting predominantly positive associations. DIN shared similar negative associations with cropland proportion, sea surface salinity, and population density, but additionally showed positive associations with upper ocean heat

content in the Bohai Sea, Yellow Sea, and northern East China Sea, suggesting that thermal stratification may play a role in surface nutrient retention. DIP exhibited extremely weak associations with all individual factors, except for distance to coast, which showed the strongest and most spatially coherent positive pattern, indicating that proximity to land-based phosphorus sources is the primary correlate of DIP distribution. Lastly, the spatial patterns of SHAP effects consistently converged in the Bohai Bay across all three pollutants, highlighting this semi-enclosed sea as a region where multiple pollutant sources converge and are retained due to limited water exchange, resulting in compounded positive associations across indicators and factors. This spatially explicit evidence supports the formulation of differentiated pollution control strategies tailored to the distinct driving mechanisms operating in different coastal regions of China.

Author Contributions: Conceptualization, H.H. (Huizhen Huang) and Z.S.; methodology, H.H. (Huizhen Huang) and Z.S.; software, H.H. (Huizhen Huang) and K.Z.; validation, H.H. (Huizhen Huang), Y.Y. and X.Z.; formal analysis, H.H. (Huizhen Huang), Z.S. and K.Z.; investigation, H.H. (Huizhen Huang), Y.Y. and C.Q.; data curation, H.H. (Huizhen Huang), Y.Y. and C.Q.; writing—original draft preparation, H.H. (Huizhen Huang) and Z.S.; writing—review and editing, H.H. (Honghao Hu), Y.Y., X.Z. and Z.S.; visualization, H.H. (Huizhen Huang); supervision, Z.S. and K.Z.; project administration, Z.S.; funding acquisition, H.H. (Huizhen Huang) and Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Zhangzhou Joint Fund for Natural Sciences (grant number ZZ2025LH29); Zhangzhou Institute of Technology Doctoral Research Start-up Fund Project in 2023 (grant number ZZYB2305); Zhangzhou Institute of Technology Horizontal Scientific Research Project (grant number HX250067).

Data Availability Statement: The final processed datasets and the R code used for statistical analysis and modeling in this study are publicly available in the Zenodo repository (accessible at: <https://doi.org/10.5281/zenodo.21187533>).

Acknowledgments: The authors would like to thank the editor and anonymous reviewers for their useful advice that helped to improve the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Atapaththu, K.S.S.; Herath, S.S.; Subramaniam, G. Challenges in coastal ecosystem sustainability: Drivers of water quality degradation and their ecological impact. *Mar. Environ. Res.* **2025**, *209*, 107194.
2. Dai, B.L.; Hua, Z.L.; Mu, F.H.; Xu, N.; He, Y.L. Offshore ecosystem health status assessment: A review. *Chin. J. Appl. Ecol.* **2013**, *24*, 1169–1176. (In Chinese)
3. Tuholske, C.; Halpern, B.S.; Blasco, G.; Villasenor, J.C.; Frazier, M.; Caylor, K. Mapping global inputs and impacts from human sewage in coastal ecosystems. *PLoS ONE.* **2021**, *16*, e0258898.
4. Sundhar, S.; Shakila, R.J.; Shalini, R.; Aanand, S.; Jayakumar, N. Spatial distribution and seasonal variation of pesticide residues in the sediment, seawater, and edible seaweeds: Environmental and human health risk assessment. *Mar. Pollut. Bull.* **2023**, *194*, 115435.
5. Wang, L.; Zheng, M.; Xu, H.; Hua, Y.; Liu, A.; Li, Y.; Fang, L.; Chen, X. Fate and ecological risks of current-use pesticides in seawater and sediment of the Yellow Sea and East China Sea. *Environ. Res.* **2022**, *207*, 112673.
6. Cai, C.; Devassy, R.P.; El-Sherbiny, M.M.; Agusti, S. Cement and oil refining industries as the predominant sources of trace metal pollution in the Red Sea: A systematic study of element concentrations in the Red Sea zooplankton. *Mar. Pollut. Bull.* **2022**, *174*, 113221.

7. Du, Z.; Qi, J.; Wu, S.; Zhang, F.; Liu, R. A spatially weighted neural network based water quality assessment method for large-scale coastal areas. *Environ. Sci. Technol.* **2021**, *55*, 2553–2563.
8. Wu, M.L.; Wang, Y.S.; Gu, J.D. Assessment for water quality by artificial neural network in Daya Bay, South China Sea. *Ecotoxicology* **2015**, *24*, 1632–1642.
9. Wu, M.; Wang, Y.; Dong, J.; Sun, F.; Wang, Y.; Hong, Y. Spatial assessment of water quality using chemometrics in the Pearl River Estuary, China. *Front. Earth Sci.* **2017**, *11*, 114–126.
10. Huang, P.; Hou, M.Y.; Sun, T.; Xu, H.; Ma, C.; Zhou, A. Sustainable groundwater management in coastal cities: Insights from groundwater potential and vulnerability using ensemble learning and knowledge-driven models. *J. Clean. Prod.* **2024**, *442*, 141152.
11. Su, Z.; Lin, L.; Xu, Z.; Chen, Y.; Yang, L.; Hu, H.; Lin, Z.; Wei, S.; Luo, S. Modeling the effects of drivers on PM2.5 in the Yangtze River Delta with geographically weighted random forest. *Remote Sens.* **2023**, *15*, 3826.
12. Sahbani, S.; Bejaoui, B.; Fathalli, A. Benabdallah, C.S.; Idriss, R.; Elhasni, K.; Ottaviani, E. TRIX-XGBoost trophic model for assessing the trophic status of Ichkeul Lake, a Ramsar sanctuary in northern Tunisia. *Mar. Pollut. Bull.* **2025**, *217*, 118148.
13. Zhu, L.; Cui, T.; Runa, A.; Pan, X.; Zhao, W.; Xiang, J.; Cao, M. Robust remote sensing retrieval of key eutrophication indicators in coastal waters based on explainable machine learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *211*, 262–280.
14. Xiao, G.; Wang, T.; Luo, Y.; Yang, D. Analysis of port pollutant emission characteristics in United States based on multiscale geographically weighted regression. *Front. Mar. Sci.* **2023**, *10*, 1131948.
15. Zhang, X.Y.; Xu, Z.H.; Fan, W.J.; Chen, K.; Yang, Z. Satellite-based estimation of spatiotemporal variation of dissolved inorganic nitrogen in the coastal waters of China. *J. Hydroecol.* **2025**, *46*, 1–9. (In Chinese)
16. Cen, H.; Han, G.; Lin, X. Deep learning-based forecasting model for chlorophyll-a response to tropical cyclones in the Western North Pacific. *Ocean Model.* **2024**, *189*, 102345.
17. Wang, H.; Chen, K.L.; Wu, Y.F. Long time-series analysis and flux estimation of dissolved inorganic nitrogen at the river–sea boundary of the Minjiang Estuary. *Environ. Monit. China* **2023**, *6*, 98–109. (In Chinese)
18. Lin, J.; Wang, P.; Wang, J.; Zhou, Y.; Zhou, X.; Yang, P.; Zhang, H.; Cai, Y.; Yang, Z. An extensive spatiotemporal water quality dataset covering four decades (1980–2022) in China. *Earth Syst. Sci. Data* **2024**, *16*, 1137–1149.
19. Helsel, D.R. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*; Wiley: New York, NY, USA, 2005.
20. *HJ 828-2017*; Water Quality—Determination of Chemical Oxygen Demand—Dichromate Method. Ministry of Ecology and Environment of China: Beijing, China, 2017.
21. *GB 17378.4-2007*; The Specification for Marine Monitoring—Part 4: Seawater Analysis. General Administration of Quality Supervision, Inspection and Quarantine of China: Beijing, China, 2007.
22. Quinn, K.J.; Ponte, R.M. High frequency barotropic ocean variability observed by GRACE and satellite altimetry. *Geophys. Res. Lett.* **2012**, *39*, L07602.
23. Piecuch, C.G.; Quinn, K.J.; Ponte, R.M. Satellite-derived interannual ocean bottom pressure variability and its relation to sea level. *Geophys. Res. Lett.* **2013**, *40*, 3106–3110.
24. Fulke, A.B.; Panigrahi, J.; Eranezhath, S.; Karthi, J.; Dora, G.U. Environmental variables and its association with faecal coliform at Madh Island beaches of megacity Mumbai, India. *Environ. Pollut.* **2024**, *341*, 122885.
25. Abdul Wahid, A.; Arunbabu, E. Multivariate analysis of water quality dynamics in a highly eutrophic reservoir: Hydrological, meteorological, and environmental contributions. *Stoch. Environ. Res. Risk Assess.* **2025**, *39*, 2974.
26. Pradhan, S.; Senthilkumar, P.; Pradhan, U.; Naik, S.; Sahoo, R.K.; Gera, A.; Panda, U.S.; Mishra, P.; Murthy, M.V.R.; Babu, P.V. Short-term response of Mandous severe cyclone on coastal water quality along southwestern Bay of Bengal, India. *Pure Appl. Geophys.* **2025**, *182*, 3708.
27. Zuo, H.; Balmaseda, M.A.; Tietsche, S.; Mogensen, K.; Mayer, M. The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. *Ocean Sci.* **2019**, *15*, 779–808.
28. Amante, C.; Eakins, B.W. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. NOAA Tech. Memo. NESDIS NGDC-24; National Geophysical Data Center, NOAA: Boulder, CO, USA, 2009.
29. Mitchell, A. *The ESRI Guide to GIS Analysis*; ESRI Press: Redlands, CA, USA, 2005; Volume 2.
30. Luo, S.; Su, Z.; Wei, S.; Zhong, Y.; Chen, Y.; Li, X.; Zhou, Y.; Liu, Y.; Wu, Z. The climate–fire–carbon nexus in tropical Asian forests: Fire behavior as a mediator and forest type-specific responses. *Forests* **2025**, *16*, 1544.
31. Rilling, J.; Tang, Y. A new p-value based multiple testing procedure for generalized linear models. *Stat. Comput.* **2025**, *35*, 69.
32. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.

33. Chen, X.; Wang, X.; Guo, J. Seasonal variability of the sea surface salinity in the East China Sea during 1990–2002. *J. Geophys. Res. Oceans* **2006**, *111*, C05007.
34. Zhao, W.L.; Yang, S.Y.; Wang, J.; Xiao, J.M.; Lu, X.X.; Lin, J.; Huang, P.; Cai, M.G. Load estimation and assessment of land-based pollution for Quanzhou Bay and their relevance to the Total Quantity Control of Pollutants Discharged into the Sea (TQCPS) Program in China. *Estuar. Coast. Shelf Sci.* **2015**, *166*, 230–239.
35. Ji, C.X.; Yang, G.P.; Chen, Y.; Zhang, P.Y. Distribution, degradation and bioavailability of dissolved organic matter in the East China Sea. *Biogeochemistry* **2019**, *142*, 229–246.
36. Wang, X.; Liu, L.; Zhao, L.; Xu, H.; Zhang, X. Assessment of dissolved heavy metals in the Laoshan Bay, China. *Mar. Pollut. Bull.* **2019**, *149*, 110608.
37. Jiang, Z.; Guo, J.; Chu, X.; Yu, Z.; Fang, X. The impact of vessel traffic flow and wind data on port water quality: Insights from the Zhanjiang Bay, China. *Water Air Soil Pollut.* **2025**, *236*, 8235.
38. Qin, C.; He, J.; Yu, C.; Sun, A.; Li, K.; Hu, H.; Hu, T.; Ye, Y.; Yu, Y.; Wang, H. Evaluating the water quality of the rice–fish co-culture pattern based on the modified NSF water quality index model. *Aquaculture* **2025**, *597*, 741931.
39. Lv, X.Z.; Yu, X.X.; Fan, D.X.; Li, Q.Y. Estimation of non-point source pollution loads caused by soil erosion in China. *J. Food Agric. Environ.* **2012**, *10*, 1045–1050.
40. Chen, B.B.; Li, S.L.; von Strandmann, P.A.E.P.; Sun, J.; Zhong, J.; Li, C.; Ma, T.T.; Xu, S.; Liu, C.Q. Ca isotope constraints on chemical weathering processes: Evidence from headwater in the Changjiang River, China. *Chem. Geol.* **2020**, *531*, 119783.
41. Zhang, M.; Lu, Q.; Liu, Y.; Zhang, Y.; Wang, D.; Hu, Q.; Cui, Z. Spatiotemporal evolution of chemical oxygen demand and its impact factors in Laizhou Bay. *Trans. Oceanol. Limnol.* **2025**, *47*, 121–132. (In Chinese)
42. Grant, C.; Ziadi, N.; Gagnon, B.; Flaten, D.; Schoenau, J. Integrating knowledge of nutrient forms and dynamics into improved nutrient management practices: A tribute to Regis Simard. *Can. J. Soil Sci.* **2009**, *89*, 133–144.
43. Grieger, S.R.; Harrison, J.A. Long-term disconnect between nutrient inputs and riverine exports in a semi-arid, agricultural watershed: Yakima River Basin 1945–2012. *J. Geophys. Res. Biogeosci.* **2021**, *126*, e2020JG006072.
44. Wang, M.; Ma, L.; Stokal, M.; Ma, W.; Liu, X.; Kroeze, C. Hotspots for nitrogen and phosphorus losses from food production in China: A county-scale analysis. *Environ. Sci. Technol.* **2018**, *52*, 5782–5791.
45. Huang, H.; Ke, Z.; Zhou, Z.; Zhou, W. Distribution of nutrients and eutrophication characteristics in the surface water around Nan'ao Island, South China Sea. *J. Mar. Sci.* **2025**, *47*, 196–207. (In Chinese)
46. Chou, W.C.; Sheu, D.D.; Chen, C.T.A.; Wen, L.S.; Yang, Y.; Wei, C.L. Transport of the South China Sea subsurface water outflow and its influence on carbon chemistry of Kuroshio waters off southeastern Taiwan. *J. Geophys. Res. Oceans* **2007**, *112*, C12008.
47. Chen, B.; Chen, Y.; Chen, H. Chemical oxygen demand inversion in typical coastal breeding areas in Fujian based on Sentinel-2 data. *Bull. Surv. Mapp.* **2025**, *4*, 75–81. (In Chinese)
48. Wen, L.; Li, Z.; Wen, X. Research on spatiotemporal distribution patterns of surface water COD in Luzhou based on GIS technology. *Water Resour. Dev. Manag.* **2025**, *7*, 74–78. (In Chinese)
49. Giri, S.; Qiu, Z. Understanding the relationship of land uses and water quality in twenty-first century: A review. *J. Environ. Manage.* **2016**, *173*, 41–48.
50. Zhang, Y.; Li, Q.; Zhao, J.; Cui, Z.; Zhou, M.; Zhu, J.; Ding, D.; Guo, F.; Liu, C.; Qu, K. Temporal and spatial distribution of COD and its source and contribution to eutrophication in the central Bohai Sea. *Prog. Fish. Sci.* **2016**, *37*, 43–48.
51. Liu, W.R.; Zhao, J.L.; Liu, Y.S.; Chen, Z.F.; Yang, Y.Y.; Zhang, Q.Q.; Ying, G.G. Biocides in the Yangtze River of China: Spatio-temporal distribution, mass load and risk assessment. *Environ. Pollut.* **2015**, *200*, 53–63.
52. Lu, Y.; Pan, D.; Luan, Y.; Ding, J.; Zhang, Z. Size fraction, behavior, and fate of trace metals along the river-estuary-coastal ocean continuum of Shandong Peninsula, northern China. *Mar. Environ. Res.* **2025**, *210*, 106948.
53. Lin, Y.J.; Gong, D.C.; Lv, S.J.; Ding, Y.Z.; Wu, G.C.; Wang, H.; Li, Y.L.; Wang, Y.J.; Zhou, L.; Wang, B.G. Observations of high levels of ozone-depleting CFC-11 at a remote mountain-top site in southern China. *Environ. Sci. Technol. Lett.* **2019**, *6*, 114–118.
54. Li, B.; Gu, W.; Jin, Y.; Ma, Y.; Bai, Y.; Sun, S.; Wang, W.; Liu, X. Distribution of chlorophyll-a and primary productivity in Yantai Sishili Bay. *Prog. Fish. Sci.* **2012**, *33*, 19–23. (In Chinese)
55. Li, X.Z.; Li, C.H.; Bai, Y.; Shi, X.Y.; Su, R.G. Composition variations and spatiotemporal dynamics of dissolved organic matters during the occurrence of green tide (*Ulva prolifera* blooms) in the Southern Yellow Sea, China. *Mar. Pollut. Bull.* **2019**, *146*, 619–630.
56. Zhai, T.; Wang, J.; Fang, Y.; Qin, Y.; Huang, L.; Chen, Y. Assessing ecological risks caused by human activities in rapid urbanization coastal areas: Towards an integrated approach to determining key areas of terrestrial-oceanic ecosystems preservation and restoration. *Sci. Total Environ.* **2020**, *708*, 134796.

57. Kim, T.H.; Kim, G.; Shen, Y.; Benner, R. Strong linkages between surface and deep-water dissolved organic matter in the East/Japan Sea. *Biogeosciences* **2017**, *14*, 2561–2570.
58. Fan, K.; Wang, X.; Liu, J.; Shao, C. Spatiotemporal variability of tropical cyclone-induced ocean heat uptake and its effect on ocean heat content. *J. Climate* **2023**, *36*, 3481–3497.
59. Zhang, H.; Xin, M.; Lin, C.; Wang, B.; Ouyang, W.; Liu, X.; He, M. Phosphorus distribution in the water and sediment of Laizhou Bay and sediment phosphorus release potential. *Sci. Total Environ.* **2022**, *846*, 157340.
60. Southwell, M.W.; Kieber, R.J.; Mead, R.N.; Brooks Avery, G.; Skrabal, S.A. Effects of sunlight on the production of dissolved organic and inorganic nutrients from resuspended sediments. *Biogeochemistry* **2010**, *98*, 115–126.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.