applied sciences

MDPI

Article

Multivariate Regression Analysis for Identifying Key Drivers of Harmful Algal Bloom in Lake Erie

Omer Mermer [1,*] and Ibrahim Demir [2,3]

[1] IIHR Hydroscience and Engineering, University of Iowa, Iowa City, IA 52242, USA
[2] Department of River-Coastal Science and Engineering, Tulane University, New Orleans, LA 70118, USA; idemir@tulane.edu
[3] ByWater Institute, Tulane University, New Orleans, LA 70118, USA
* Correspondence: omer-mermer@uiowa.edu

Abstract: Harmful Algal Blooms (HABs), predominantly driven by cyanobacteria, pose significant risks to water quality, public health, and aquatic ecosystems. Lake Erie, particularly its western basin, has been severely impacted by HABs, largely due to nutrient pollution and climatic changes. This study aims to identify key physical, chemical, and biological drivers influencing HABs using a multivariate regression analysis. Water quality data, collected from multiple monitoring stations in Lake Erie from 2013 to 2020, were analyzed to develop predictive models for chlorophyll-a (Chl-a) and total suspended solids (TSS). The correlation analysis revealed that particulate organic nitrogen, turbidity, and particulate organic carbon were the most influential variables for predicting Chl-a and TSS concentrations. Two regression models were developed, achieving high accuracy with R^2 values of 0.973 for Chl-a and 0.958 for TSS. This study demonstrates the robustness of multivariate regression techniques in identifying significant HAB drivers, providing a framework applicable to other aquatic systems. These findings will contribute to better HAB prediction and management strategies, ultimately helping to protect water resources and public health.

Keywords: HAB; Chlorophyll-a; total suspended solids; water quality parameters; linear regression model; Pearson's correlation coefficient; ANOVA test; relative importance

check for updates

Academic Editor: Ricardo Castedo

Received: 31 March 2025
Revised: 22 April 2025
Accepted: 24 April 2025
Published: 26 April 2025

Citation: Mermer, O.; Demir, I. Multivariate Regression Analysis for Identifying Key Drivers of Harmful Algal Bloom in Lake Erie. *Appl. Sci.* **2025**, *15*, 4824. https://doi.org/10.3390/app15094824

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Environmental contamination has contributed to a significant increase in cyanobacterial biomass (i.e., algae blooms) in water bodies, severely affecting water quality [1]. The term "bloom" refers to the rapid growth of blue-green algae, or cyanobacteria, which can produce harmful toxins [2]. The emergence and proliferation of Harmful Algal Blooms (HABs), primarily driven by cyanobacteria, has become a critical environmental concern worldwide. These blooms degrade water quality, threaten public health, and disrupt aquatic ecosystems. Key factors fueling the rise in HAB incidents include nutrient pollution, particularly from agricultural runoff and industrial waste [3], as well as climatic changes, such as rising water temperatures and shifts in water quality [4,5]. HABs are known for producing hazardous toxins, undermining the recreational and aesthetic value of waterways, and complicating efforts to provide clean drinking water [6]. The recent surge in HAB occurrences has been linked to population growth, intensified agricultural practices [7,8], increasing pollution levels, and climate change [9]. This trend underscores the urgent need for improved HAB monitoring, estimation, modeling, and prediction techniques to safeguard water resources and public health [10–13].

Lake Erie, as part of the Great Lakes system, provides a compelling case study for investigating HABs. The Great Lakes represent the largest and most biodiverse freshwater system on Earth [14,15]. With both industrial and agricultural regions in its basin, Lake Erie is the shallowest and smallest of the Great Lakes by volume, yet it holds ecological, cultural, and economic importance for approximately 12.5 million residents within its watershed. Lake Erie supports commercial and traditional fisheries, extensive freight transport, and a robust recreation and tourism industry [16]. Its western basin is particularly prone to nutrient overload, primarily due to its geographical setting [17]. Since 2002, chlorophyll-a (Chl-a) concentrations, a widely accepted indicator of eutrophication and HABs, have risen to unprecedented levels [17,18]. Humans face exposure to HABs through ingestion, drinking water, and recreational activities [19,20]. Therefore, predicting HAB occurrences is essential for minimizing health, economic, and environmental risks. Identifying the key factors driving these blooms is crucial for implementing effective mitigation strategies [21].

HAB formation typically results from the interplay of various factors that foster favorable growth conditions [22]. Eutrophication, poor water quality, especially elevated nitrogen and phosphorus levels, and climate change are among the primary contributors [23,24]. While past studies have examined individual drivers such as nutrients, land use, or climate, few have explored the complex interactions between physical, chemical, and biological factors [25–27]. Previous studies, such as Hushchyna et al. [28], have highlighted key nutrient drivers like total phosphorus (TP) and iron in predicting cyanobacterial bloom intensity in Lake Torment, Nova Scotia, emphasizing the broader significance of nutrient management in mitigating HABs across freshwater systems. These approaches often rely on limited data, which can oversimplify predictions and introduce potential inaccuracies. Therefore, understanding the multifaceted drivers of HABs is critical for effective monitoring and prevention.

Traditional HAB monitoring techniques, including laboratory analysis of chlorophyll-a, cyanobacteria, and various algal toxins, are labor-intensive and require specialized expertise [29–31]. Remote sensing technologies, such as satellite and Unmanned Aerial Vehicle (UAV) data, offer valuable spatial insights into understanding environmental drivers [32,33], including bloom dynamics [34–36]. However, the relationship between chlorophyll-a levels and algal toxicity is complex, varying by location and environmental conditions [37,38]. Furthermore, higher chlorophyll-a concentrations do not always signify high toxin levels but may indicate a higher probability of exceeding certain thresholds [38]. Therefore, understanding the main drivers of HABs remains pivotal for accurate prediction and prevention efforts.

Various prediction models have been developed to address HAB dynamics. Physical process-based models, such as the Environmental Fluid Dynamic Code (EFDC) [39], Water Quality Simulation 2000 (QUAL2K) [40], and Water Quality Analysis Simulation Program (WASP) [41], rely on detailed physical and biochemical processes but face challenges in handling spatial data and high costs [42–45]. While these models have high prediction accuracy with complete data, constructing perfect data, particularly with spatial resolution, is costly and involves practical limitations. Statistical models that correlate physicochemical and meteorological variables often struggle with capturing nonlinear relationships [46,47], limiting their predictive accuracy [48]. Both approaches contribute valuable insights but face limitations in predicting HAB dynamics under varied conditions [49].

To address these challenges, data-driven models have emerged as promising alternatives for predicting HABs. These models, increasingly applied in hydrology, water resources, and environmental management, can uncover complex relationships without the need for explicit mathematical modeling of unknown processes [50,51]. Data-driven approaches [52,53] offer a significant advantage in analyzing and predicting HAB occur-

rences, providing more accurate and reliable results, and support informed decisions for the public through visualization and communication systems [54,55].

The objective of this study is to quantitatively identify the key physical, chemical, and biological drivers of HABs using multivariate regression analysis. Specifically, we aim to (1) assess the relative importance of explanatory variables, (2) identify the major drivers of HAB dynamics, and (3) quantify the relationship between water quality variables and HAB formation. Statistical methods such as correlation analysis, multivariate regression analysis, and the ANOVA F-test were applied to our dataset. In addition, multivariate regression analysis was also conducted for total suspended solids (TSS), which is an important indicator for monitoring water quality and measuring the degree of water pollution, to further demonstrate the robustness of our approach. The HAB analysis framework developed in this study is designed for broad application to lakes, rivers, and coastal areas. We anticipate that this framework will serve as a valuable tool for scientists and stakeholders, offering practical guidance for understanding and mitigating HAB risks globally.

This paper is structured as follows: Section 2 describes the methodology, including data collection and correlation and multivariate regression analysis. Section 3 presents the results and discussion, highlighting the key drivers for HAB. Section 4 concludes with insights into the implications of the findings and suggestions for future research.

2. Materials and Methods

Figure 1 outlines the workflow for the proposed study, starting with dataset preparation, where water quality data undergoes preprocessing to address missing values, remove duplicate records, and ensure data consistency. In the Correlation Analysis phase, the relationships between the water quality variables and the target variables (Chl-a and TSS) are examined using Pearson's correlation coefficient. Additionally, the relative importance (RI) index is employed to determine the influence of each predictor variable, while bivariate scatter plots visualize the linear relationships. This analysis identifies the key water quality parameters that exhibit statistically significant correlations with the target variables. The third stage involves Multivariate Regression Analysis, where two separate models are developed: Model 1, which focuses on predicting Chl-a concentrations, and Model 2, which targets the prediction of TSS levels. These models aim to uncover the functional relationships between the predictors and the target variables, considering the complexity and interdependence of the environmental factors. All statistical analyses and visualizations were performed using Python (v3.11.7) with the statsmodels, scikit-learn, and matplotlib packages.

2.1. Study Area and Dataset Preprocessing

As the nutrient load and HAB occurrence in Lake Erie take place mostly in the western part, our study focused on the western basin of Lake Erie (shown in Figure 2), which encompasses the western part of the lake to Point Pelee, ON, Canada, and Cedar Point, OH, USA. In this study, we used water quality data collected by the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) from multiple monitoring stations (Figure 2) on the US side of the western part of Lake Erie. The data span the years 2013 to 2020, with a near-daily resolution during the bloom-prone months (typically May through September) [17]. A comprehensive description of the sampling methodology, including sampling frequency, station characteristics, equipment setup, calibration routines, and deployment strategies, can be found in Boegehold et al. [17], which details NOAA's HAB monitoring framework in western Lake Erie. We specifically focused on monitoring stations closest to Maumee River inflow (i.e., WE06 and WE09),

which reflect the various nutrient and sediment input into the western basin of Lake Erie as well as representing the areas that are prone to HAB consistently compared to other stations [17,18]. All data preprocessing was performed by the authors and included the removal of duplicate entries, handling of missing values, and consistency checks to ensure high-quality inputs for statistical analysis. In terms of missing data handling, records with missing values in either the target variables (Chl-a, TSS) or in any of the key predictor variables selected for regression modeling were excluded from the analysis. Due to the high temporal resolution of the NOAA monitoring dataset, these omissions represented a small fraction of the total data and did not introduce significant bias. We opted not to perform imputation to avoid introducing artificial variance or assumptions into the regression models.

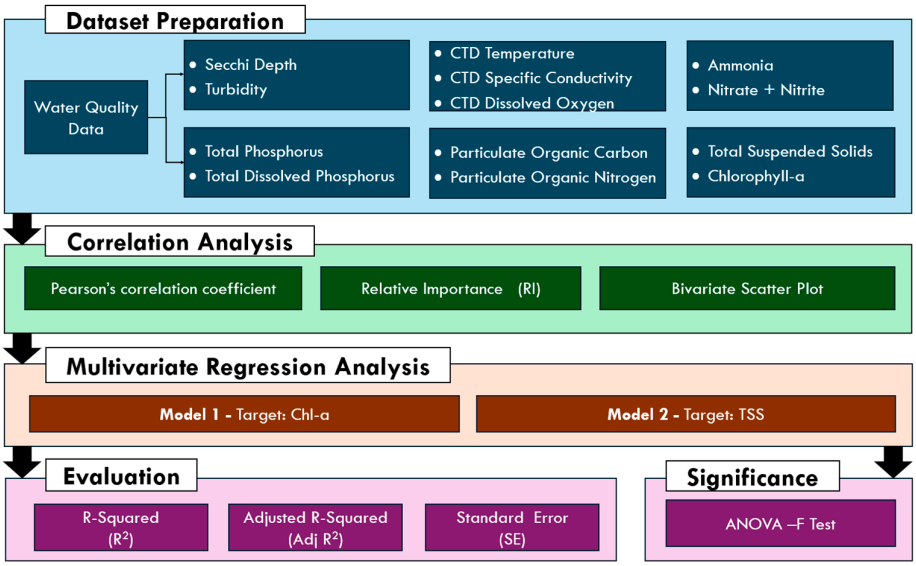


Figure 1. Workflow of the proposed study for analyzing the key drivers of HABs.

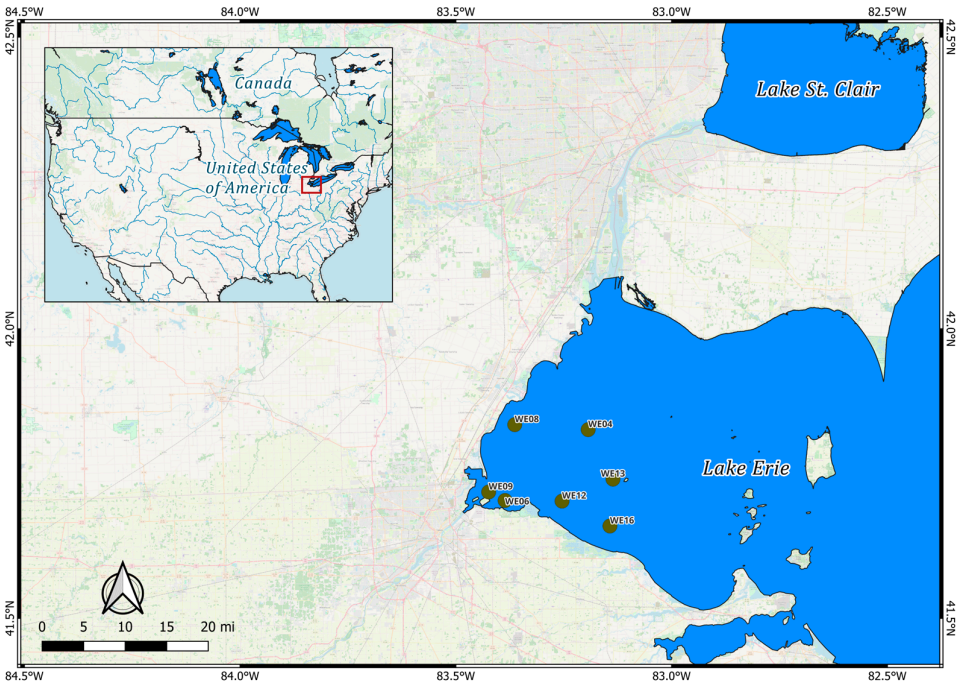


Figure 2. Location and description of western Lake Erie water quality monitoring stations operated by NOAA's Great Lakes Environmental Research Laboratory. Retrieved from https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/rtMonSQL.php (accessed on 16 October 2024).

Based on the available data, we selected physical, chemical, and biological variables, shown in Table 1. Physical variables include Secchi Depth (SD), CTD Temperature (T), CTD Specific Conductivity (Cond), and Turbidity (Turb). Chemical drivers also include CTD Dissolved Oxygen (DO), Total Phosphorus (TP), Total Dissolved Phosphorus (TDP), Ammonia (A), Nitrate + Nitrite (NOx), Particulate Organic Carbon (POC), Particulate Organic Nitrogen (PON), and Total Suspended Solids (TSS). Chlorophyll-a (Chl-a) is also categorized under the biological variables.

Table 1. Summary of the variables used in this study.

Variable	Unit	Definition
Secchi Depth	m	Penetration depth of sunlight through the water
CTD Temperature	°C	Water temperature at site
CTD Specific Conductivity	µS/cm	Conductivity value of water at site
CTD Dissolved Oxygen	mg/L	Concentration of dissolved oxygen at site
Turbidity	NTU	Cloudiness of a fluid caused by suspended solids
Total Phosphorus	µg/L	Concentration of the sum of all phosphorus compounds that occur in various forms at site
Total Dissolved Phosphorus	µg/L	Concentration of the portion of phosphorus that is dissolved at site
Ammonia	µg/L	Concentration of Ammonia at site
Nitrate + Nitrite	mg N/L	Concentration of NO_x at site
Particulate Organic Carbon	mg/L	Concentration of organic carbon particles suspended in water at site
Particulate Organic Nitrogen	mg/L	Concentration of organic nitrogen particles suspended in water at site
Total Suspended Solids	mg/L	Concentration of both organic and inorganic particles suspended in water at site
Chlorophyll-a	µg/L	Indicator of HABs

2.2. Correlation Analysis

In this work, Pearson's correlation coefficient, which is a pivotal statistical technique, was used to analyze the correlation between two variables. The Pearson correlation coefficient method is aimed at assessing linear variable correlations. The correlation coefficient is typically denoted by the symbol r whose values range between -1 and 1, where 1 indicates strong positive correlation, -1 indicates strong negative correlation, and a value near 0 suggests a weak or nonexistent relationship [56]. r can be calculated using Equation (1):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (1)$$

where r is the Pearson's correlation coefficient, n is the sample size, x_i and y_i are the ith sample values, and $\overline{x}$ and $\overline{y}$ are the mean values of x and y. In addition, an absolute value of r more than 0.8 means a strong correlation, <0.2 means a weak correlation, and between 0.2 and 0.8 indicates a correlation [57].

2.3. Multivariate Regression Analysis

Multivariate analysis is essential for categorizing environmental drivers (i.e., HABs) with similar traits and summarizing related multivariate patterns, offering valuable insights for creating targeted mitigation strategies [58,59]. Multivariate regression analysis was performed to develop a regression model between the dependent variable and different independent variables. The Chl-a and TSS were set as the dependent variables for our study. To avoid the influence of collinearity on the regression analysis, independent variables were selected by comparing the relative importance values. A simple correlation analysis was performed to estimate the correlations between the dependent variable and independent variables. Then, multivariate regression analysis was performed using the least squares

method. When there are n independent variables (X_i), the dependent variable (Y) can be described under the form of the below equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \epsilon \quad (2)$$

From Equation (2), it can be seen that the regression coefficient, or slope, b_i (unbiased estimate), represents the change in Y per unit change in the (X_i) variable after the adjustment for simultaneous linear change; and the y-intercept, b_0, also called the multiregression constant, standing for the y value where the regression line crosses the y-axis. Hence, it is the value of y when the value of x is equal to 0. The last parameter ϵ in Equation (2) represents the residuals (error term). Equation (2) can be very helpful in predicting the value of the dependent variable (Y) from the given value of the independent variables (X_i). It also may predict Y from the outer given ranges of (X_i), but such extrapolation is not highly recommended [60]. We have also herein the ANOVA-F test in our regression model to nullify our hypothesis that there is no relationship between the independent variable (X) and dependent (Y) variable, i.e., all regression coefficients equal to zero ($b_1 = b_2 = \ldots = b_n = 0$). From the ANOVA-F test, the significant p-value (<0.05 at 95% confidence interval) suggests that the relationship between (X_i) and Y is crucial. The independent variables (X_i) can reliably predict the dependent variable Y.

2.4. Coefficients of Determination

Coefficients of determination (R^2) indicate how well the predictive value explains the measured value. In this work, measured and predicted Chl-a (and TSS) concentrations were taken as the dependent and independent variables, respectively, and the R^2 was determined by applying linear regression analysis. The R^2 ranges from 0 to 1; the closer the value is to 1, the better the independent variable explains the dependent variable, meaning the higher the prediction accuracy. The formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (3)$$

where y_i is the ith actual value, $\hat{y}$ is the ith predicted value of the dependent variable, and $\overline{y}$ is the mean of y_i. The adjusted R^2, which is defined as the corrected value of R^2 for sample size and regression coefficients, is a better parameter than R^2 itself. The adjusted R^2 is always less than R^2. A higher adjusted R^2 generally represents a better model but is not always correct and should be used with caution to assess the model. There is no cutoff point of R^2 for the appropriate model selection. R^2 should be evaluated based on the field data types, data transformations, or subject area decisions [61].

3. Results and Discussions

3.1. Data Statistics

Table 2 shows a summary of the descriptive statistics of the dataset after preprocessing. In the table, POC and PON levels varied widely as well, with maximums far exceeding the average, suggesting the presence of organic matter in different concentrations throughout the samples. On the other hand, TSS and Chl-a also showed large ranges in value, pointing to varied conditions in the sampled water bodies. The standard deviation for each variable suggests the extent of variation in the measurements, with some variables like turbidity and total phosphorus showing very high variability. It is noticed that we have in total 13 parameters, and we distinguish these parameters into two different targets: Chl-a and TSS (as one target) and the predictors (the rest of the parameters).

Table 2. Statistical description of the dataset used in this study.

Variables	Min	Max	Mean	Std. Dev.
SD	0	5.3	0.796	0.694
T	10.1	29.7	22.417	3.651
Cond	19.9	583.3	337.586	67.828
DO	4.2	13.0	7.478	1.217
Turb	0.95	1148.0	29.599	78.295
TP	14.87	2482.2	119.144	181.173
TDP	2.67	273.6	30.909	34.865
A	0.04	561.6	39.822	56.930
NOx	0	9.5	1.308	1.676
POC	0.15	219.3	3.946	15.381
PON	0.03	40.9	0.677	2.759
TSS	1.25	540.8	25.489	44.275
Chl-a	0.71	6784.0	61.232	347.307

3.2. Correlation and Relative Importance

Table 3 presents the relative importance (RI) values for each predictor variable in relation to Chl-a and TSS. For Chl-a, PON, Turbidity, and POC emerge as the most influential variables, contributing 24.26%, 22.77%, and 18.39% to the model, respectively. These factors play a critical role in explaining the variability of Chl-a concentrations in the water. Other significant contributors include TSS and TP, accounting for 13.00% and 10.01%, respectively. However, the remaining variables, such as SD, DO, and T, provide minimal contributions, and variables like Cond and NOx contribute even less.

Table 3. Relative importance (RI) value for two considered targets.

	RI for Chl-a (100%)	RI for TSS (100%)
PON	24.26	28.06
Turb	22.77	35.57
POC	18.39	24.12
TSS	13.00	—
TP	10.01	6.20
SD	3.14	0.34
DO	2.44	0.00
T	2.19	0.15
TDP	2.01	0.12
A	1.25	0.11
Cond	0.50	0.00
NOx	0.05	0.01
Chl-a	—	5.34

Table 4 presents the Pearson correlation coefficients between all target variables and regressors. Pearson correlation coefficients indicate the strength and direction of the linear relationship between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value close to 0 suggests no correlation. This correlation matrix provides important insights into the relationships between environmental variables and the two target variables, Chl-a and TSS. The strongest drivers for Chl-a are PON, Turbidity, and POC, while TSS is primarily influenced by Turbidity, PON, and POC.

Table 4. Pearson's correlation coefficients for all target variables and regressors.

	SD	T	Cond	DO	Turb	TP	TDP	A	NOx	POC	PON	TSS	Chl-a
SD	1.00	0.12	−0.13	−0.05	−0.23	−0.27	−0.15	−0.06	−0.01	−0.12	−0.11	−0.35	−0.06
T	0.12	1.00	−0.01	−0.32	−0.07	−0.06	−0.06	−0.15	0.06	0.06	0.06	−0.16	0.07
Cond	−0.13	−0.01	1.00	−0.15	−0.04	0.09	0.40	0.39	0.50	−0.06	−0.06	−0.03	−0.05
DO	−0.05	−0.32	−0.15	1.00	0.10	0.03	−0.31	−0.32	−0.23	0.16	0.16	0.06	0.19
Turb	−0.23	−0.07	−0.04	0.10	1.00	0.88	0.11	0.06	0.05	0.89	0.91	0.93	0.75
TP	−0.27	−0.06	0.09	0.03	0.88	1.00	0.28	0.14	0.19	0.76	0.78	0.83	0.69
TDP	−0.15	−0.06	0.40	−0.31	0.11	0.28	1.00	0.48	0.61	−0.08	−0.08	0.17	−0.06
A	−0.06	−0.15	0.39	−0.32	0.06	0.14	0.48	1.00	0.46	−0.09	−0.09	0.12	−0.08
NOx	−0.01	0.06	0.50	−0.23	0.05	0.19	0.61	0.46	1.00	−0.09	−0.08	0.09	−0.07
POC	−0.12	0.06	−0.06	0.16	0.89	0.76	−0.08	−0.09	−0.09	1.00	0.99	0.78	0.71
PON	−0.11	0.06	−0.06	0.16	0.91	0.78	−0.08	−0.09	−0.08	0.99	1.00	0.76	0.79
TSS	−0.35	−0.16	−0.03	0.06	0.93	0.83	0.17	0.12	0.09	0.78	0.76	1.00	0.49
Chl-a	−0.06	0.07	−0.05	0.19	0.75	0.69	−0.06	−0.08	−0.07	0.71	0.79	0.49	1.00

Figure 3 also shows a bivariate scatter plot indicating pairwise relationships between seven selected regressors, SD, Turb, TP, POC, PON, TSS, and Chl-a, for this analysis. Clear positive linear relationships are observed between Turbidity and TSS, as well as between Turbidity, POC, PON, and Chl-a, suggesting that these variables are important drivers of both suspended solids and algal bloom concentration. TP also shows moderate positive associations with both turbidity and Chl-a, reinforcing the role of phosphorus in contributing to water turbidity and algal growth.

To assess multicollinearity among predictor variables, we calculated the variance inflation factor (VIF) values for all predictors related to the Chl-a and TSS models. While most variables had VIFs below 5, a few (e.g., PON, POC, Turbidity) exhibited VIF values above the commonly accepted thresholds (VIF > 10). Nevertheless, they were retained in the final models due to their strong predictive performance and ecological importance in driving HABs. Although multicollinearity can affect the precision of individual coefficient estimates, it does not necessarily impair model validity for predictive purposes, especially in environmental applications where predictor variables are often interrelated [62–64]. We therefore prioritized model robustness and ecological relevance over coefficient orthogonality, consistent with best practices in applied regression modeling. As the focus of this study is on prediction and variable importance, rather than precise inference from individual coefficients, the inclusion of correlated variables was deemed appropriate for this applied modeling context.

3.3. Multivariate Regression Model Performance

In this study, the objective is to deal with multivariate regression showing the association between dependent and independent variables. With our data, after evaluating the relative importance and correlation analysis, two regression models were determined as below for the prediction Chl-a and TSS, including the following:

$$\textbf{Model 1: } \text{Chl-a} = 0.113\ \text{Tur} - 0.035\ \text{TP} - 3.453\ \text{POC} + 4.115\ \text{PON} - 0.03\ \text{TSS} + 0.004 \quad (4)$$

$$\textbf{Model 2: } \text{TSS} = 1.477\ \text{Tur} + 0.196\ \text{TP} + 2.125\ \text{POC} - 2.705\ \text{PON} - 0.126\ \text{Chl-a} + 0.007 \quad (5)$$

In Table 5, the multivariate correlation coefficient (r) provides insight into the strength and direction of the relationship between independent and dependent variables in our regression models. For Model 1, the r-value is 0.986, indicating a very strong positive correlation, which suggests a high-quality prediction for the dependent variable, Chl-a. Similarly, in Model 2, the r-value is 0.979, also signifying a strong predictive relationship,

this time for the dependent variable TSS. The coefficients of determination, represented by R^2 and adjusted R^2, further emphasize the model's effectiveness in explaining the variability in the outcome variable. In Model 1, R^2 is 0.973, meaning 97.3% of the variance in Chl-a can be attributed to the independent variables in the regression model. The adjusted R^2, which accounts for the number of predictors in the model, remains at 0.973, confirming that the predictors explain the vast majority of variability without overfitting.

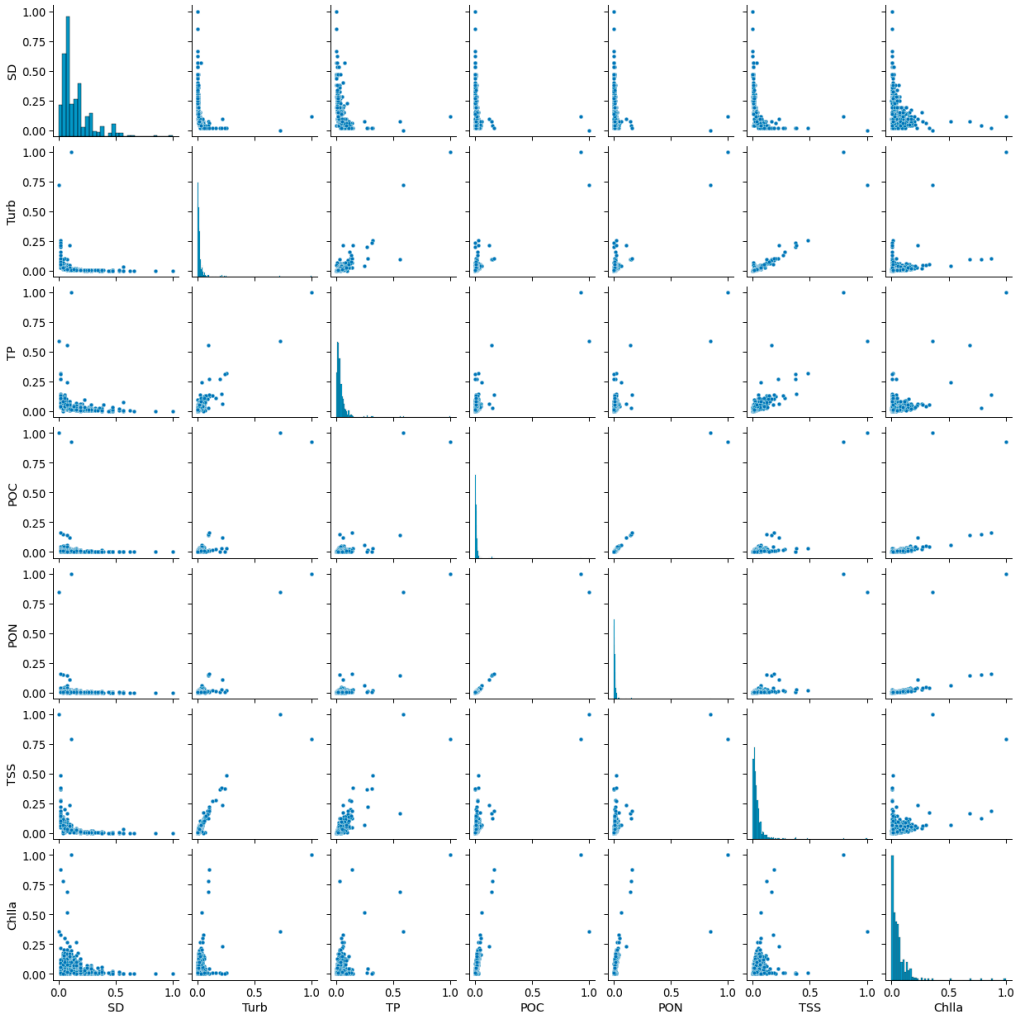


Figure 3. Bivariate scatter plots for seven selected regressors.

Table 5. Model summary.

	r	R^2	Adjusted R^2	Std. Error (SE)
Model 1	0.986	0.973	0.972	0.008
Model 2	0.979	0.958	0.957	0.016

Model 2 follows a similar trend, with R^2 at 0.958, indicating that 95.8% of the variability in TSS is explained by the independent variables. The adjusted R^2 for Model 2 also mirrors this value, reflecting the robustness of the model. The standard error (SE) of the estimate provides a measure of the average distance that the observed values fall from the regression line. For Model 1, the SE is 0.008, indicating a very small error in the predictions of Chl-a. In Model 2, the SE is slightly higher at 0.016, which still reflects a reasonably accurate prediction for TSS. In both models, the low SE values reinforce the precision and reliability of the regression models, indicating minimal deviation between the observed and predicted values.

In Table 6, the ANOVA table provides crucial information regarding the statistical significance of our regression models. The table presents two primary components: the regression and residual sum of squares, which are used to evaluate how well the independent variables explain the variance in the dependent variable. In Model 1, which predicts Chl-a, the F-ratio of 3250.46 is substantially larger than 1, indicating a highly significant regression model. The F-ratio represents the ratio of the mean square for the regression to the mean square for the residuals, which compares the variance explained by the model to the variance that remains unexplained. A high F-ratio suggests that the model explains a considerable amount of variation in the dependent variable. The associated *p*-value is reported as less than 1×10^{-4}, confirming that this result is statistically significant and that the likelihood of these findings occurring by chance is extremely low.

Table 6. ANOVA table for statistical significance.

		df	SS—Sum of Square	MS—Mean Squares	F-Ratio	*p*-Value
Model 1						
	Regression	5	0.989	1.98×10^{-1}	3250.46	$<1 \times 10^{-4}$
	Residual	453	0.028	6.09×10^{-5}		
	Total	458	1.017			
Model 2						
	Regression	5	2.598	5.20×10^{-1}	2067.33	$<1 \times 10^{-4}$
	Residual	453	0.114	2.51×10^{-4}		
	Total	458	2.712			

Therefore, we can conclude that the independent variables used in Model 1 are strong predictors of Chl-a, and the model is a good fit for the data. Similarly, for Model 2, which predicts TSS, the F-ratio is 2067.33, again significantly larger than 1. This high F-value suggests that the independent variables in this model also explain a substantial amount of the variability in TSS. The *p*-value for this model is also less than 1×10^{-4}, reinforcing the statistical significance of the model. This indicates that the relationship between the independent variables and TSS is robust, and the model fits the data well. Overall, the results presented in Table 6 demonstrate that both regression models are statistically significant, meaning that the independent variables are highly effective in predicting the respective dependent variables, Chl-a and TSS.

Table 7 presents the estimated coefficients for the multivariate regression models predicting Chl-a in Model 1 and TSS in Model 2. These unstandardized coefficients represent the direct impact of each independent variable on the dependent variable, holding all other predictors constant. Each coefficient in the regression equation explains how much the dependent variable is expected to increase or decrease with a one-unit change in the independent variable. For Model 1, the regression equation is as follows:

$$\textbf{Chl-a} = 0.113\ \text{Turb} - 0.035\ \text{TP} - 3.453\ \text{POC} + 4.115\ \text{PON} - 0.03\ \text{TSS} + 0.004$$

This equation allows for the prediction of Chl-a based on the values of five independent variables: Turb, TP, POC, PON, and TSS. For Model 2, the regression equation is as follows:

$$\textbf{TSS} = 1.477\ \text{Turb} + 0.196\ \text{TP} + 2.125\ \text{POC} - 2.705\ \text{PON} - 0.126\ \text{Chl-a} + 0.007$$

This equation will help to predict the TSS from the given value of three independent variables (Turb, TP, POC, PON, and Chl-a). The *p*-values associated with each variable provide insight into the statistical significance of these coefficients. *p*-values shown in

Table 6 are different from 0, and most of them are less than 0.05 (except TSS for model 1 and Chl-a for model 2), indicating that these variables have a meaningful effect on target variables. The standard error (SE) values in the table measure the variability of the coefficients. Although the *p*-value for TSS in Model 1 and for Chl-a in Model 2 exceeds the 0.05 threshold, the variable was retained due to its ecological significance as a proxy for nutrient-bound particulates that influence bloom formation. Smaller SE values indicate that the estimates are more precise. In Model 1, all independent variables have small SEs, suggesting stable estimates. In Model 2, although most parameters have small SEs, POC and PON have slightly larger SEs, indicating more variability in their estimates.

Table 7. Estimates of coefficients for multivariate regression models.

	Unstandardized Coefficients	Standard Error (SE)	t	*p*-Value
Model 1				
Constant	0.004	0.000	7.906	<0.0001
Turb	0.113	0.038	2.963	0.0032
TP	−0.035	0.012	−2.837	0.0048
POC	−3.453	0.076	−45.688	<0.0001
PON	4.115	0.091	45.380	<0.0001
TSS	−0.030	0.023	−1.319	0.1879
Model 2				
Constant	0.007	0.001	6.583	<0.0001
Turb	1.477	0.037	40.445	<0.0001
TP	0.196	0.023	8.392	<0.0001
POC	2.125	0.350	6.077	<0.0001
PON	−2.705	0.415	−6.521	<0.0001
Chl-a	−0.126	0.095	−1.319	0.1879

To verify the assumptions of linear regression, we conducted visual diagnostic methods for both residual normality and homoscedasticity. These included Q-Q plots for assessing the distribution of residuals and residuals vs. fitted value plots for detecting potential heteroscedasticity. Visual diagnostics are commonly accepted in applied environmental modeling as effective tools for evaluating model adequacy, particularly when dealing with complex or large datasets where slight departures from assumptions may not substantially impact model interpretation [61,65,66]. This approach is especially prevalent in environmental sciences, where residual patterns often reflect natural variability, measurement noise, or sampling limitations rather than systemic modeling flaws. Prior studies in water quality modeling, hydrology, and ecological regression have recommended visual inspection as a practical and informative method for validating model assumptions [67,68]. In this study, the Q-Q plots for both models demonstrated approximate normality, while the residuals were symmetrically distributed around zero with no clear patterns, supporting the assumption of constant variance. These results reinforce the robustness of our multivariate regression models under real-world environmental data conditions.

3.4. Discussions

This study highlights the critical role of particulate organic nitrogen (PON), turbidity, and particulate organic carbon (POC) in driving Harmful Algal Bloom (HAB) dynamics in Lake Erie. The multivariate regression models developed to predict chlorophyll-a (Chl-a) and total suspended solids (TSS) demonstrated strong predictive performance (R^2 = 0.973

for Chl-a and 0.958 for TSS), indicating that these environmental variables account for most of the variance in bloom-related water quality parameters.

However, caution must be taken in interpreting these relationships as causal. While several predictors showed strong statistical associations with the response variables, high correlation does not imply direct causation. Variables such as turbidity, PON, and POC likely act as recording indicators, reflecting an already developing bloom with increased organic matter and microplankton biomass. This interpretation is consistent with the current ecological literature on bloom progression and detection [69,70]. Total phosphorus (TP), on the other hand, can be regarded as a causal driver, as it is a well-known trigger for cyanobacterial growth [71,72] and can be actively managed through watershed-scale interventions (e.g., improved wastewater treatment, erosion control). We recommend prioritizing TP reduction in long-term nutrient management strategies [73]. Conversely, real-time increases in turbidity, PON, and POC should be used as triggers for short-term response actions, such as beach closures, public health notifications, or intensified field monitoring.

Given the temporal resolution of the data and the nature of the predictors, the developed models are best suited for short-term forecasting, typically within a 1–2 week horizon. This aligns with the early warning needs of local environmental agencies and allows for targeted intervention during bloom initiation.

Despite these promising results, it is important to acknowledge certain limitations. First, multicollinearity was present among some predictors (e.g., PON, POC, Turbidity), as indicated by high variance inflation factor (VIF) values. These variables were retained based on their ecological relevance and predictive performance. While multicollinearity may inflate coefficient variance, it does not necessarily compromise predictive validity in applied environmental modeling [62–64]. Second, the spatial scope of the data was limited to NOAA stations in the western basin. While the model may be transferable to other eutrophic lakes, regional hydrological differences—such as depth, stratification, or flow—may affect its generalizability and require local calibration. For instance, the shallow morphology and high sediment resuspension of western Lake Erie may amplify the effects of turbidity compared to deeper lakes [74,75]. Third, key hydrometeorological factors, such as wind speed, precipitation, precipitation, and temperature anomalies, are known to influence nutrient resuspension and bloom dynamics [76]. These variables were not included here due to limited integration with in situ water quality datasets. Their inclusion in future work could improve seasonal robustness and help model short-term variability. Lastly, while multivariate linear regression yielded high accuracy and interpretability, the complex, nonlinear nature of HAB processes may be better captured by advanced machine learning or hybrid modeling frameworks [77,78]. Prior research has demonstrated that the emerging approaches using remote sensing data, such as deep learning-based tools for detecting algal bloom via optical and SAR sensors [79] or RGB imaging for turbidity classification [80], offer greater flexibility for spatially and temporally dynamic environments. Addressing these limitations through the integration of broader datasets and advanced hybrid modeling techniques will be essential for scaling this framework to broader geographic areas and improving real-time HAB prediction.

4. Conclusions

This study presents a data-driven framework for identifying and quantifying key environmental drivers of Harmful Algal Blooms (HABs) in Lake Erie using multivariate regression models. Based on seven years of in situ water quality data, we developed predictive models for chlorophyll-a (Chl-a) and total suspended solids (TSS), achieving strong performance with R^2 values of 0.973 and 0.958, respectively.

Our analysis reveals that PON, turbidity, and POC are influential in predicting bloom intensity. However, it is critical to differentiate between causal drivers, such as total phosphorus (TP), and diagnostic indicators, such as turbidity and PON. While TP can be managed through watershed interventions, diagnostic indicators provide real-time feedback that a bloom is underway and can support immediate operational responses. This dual understanding informs both strategic prevention and tactical mitigation.

The proposed regression framework can support short-term forecasting (1–2 weeks) and can help guide timely decisions, such as issuing advisories, closing recreational areas, or increasing sampling efforts. Although the model is developed for the western basin of Lake Erie, the underlying methodology may be adapted for use in other nutrient-impacted freshwater systems.

While this study provides a valuable framework, there are several opportunities to address in future work. The current models rely on linear relationships and region-specific data. Incorporating meteorological, hydrological, and land-use variables, as well as satellite-derived indices and real-time monitoring platforms, may enhance both spatial transferability and early detection capacity. Moreover, integrating machine learning methods—particularly hybrid models that combine regression with deep learning—can improve the representation of complex, nonlinear interactions and enable dynamic forecasting across diverse hydroclimatic regimes.

In summary, this research offers a practical and interpretable model for HAB prediction, reinforces the importance of phosphorus reduction, and highlights actionable indicators for timely intervention. These insights provide a valuable decision-support tool for environmental managers, public health officials, and water policy makers tasked with mitigating the risks and impacts of HABs in freshwater ecosystems.

Author Contributions: O.M.: conceptualization, methodology, validation, data curation, visualization, formal analysis, and writing—original draft preparation; I.D.: writing—review and editing, validation, supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are available in the NOAA National Centers for Environmental Information (NCEI) data repository at https://doi.org/10.25921/11da-3x54 (accessed on 15 September 2023). Detailed descriptions of these data were also presented in Boegehold et al. [17].

Acknowledgments: O.M. would like to give special thanks to the Next Generation Internet Transatlantic Fellowship Program for their generous support and funding, which has been instrumental in the completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, Z.; Kamanmalek, S.; Alamdari, N.; Nikoo, M.R. Comprehensive Insights into Harmful Algal Blooms: A Review of Chemical, Physical, Biological, and Climatological Influencers with Predictive Modeling Approaches. *J. Environ. Eng.* **2024**, *150*, 03124002. [CrossRef]
2. Carmichael, W.W.; Falconer, I.R. Diseases related to freshwater blue-green algal toxins, and control measures. In *Algal Toxins in Seafood and Drinking Water*; Falconer, I.R., Ed.; Academic Press: London, UK, 1993; pp. 187–209.
3. Bayar, S.; Demir, I.; Engin, G.O. Modeling leaching behavior of solidified wastes using back-propagation neural networks. *Ecotoxicol. Environ. Saf.* **2009**, *72*, 843–850. [CrossRef]

4. Paerl, H.W.; Paul, V.J. Climate change: Links to global expansion of harmful cyanobacteria. *Water Res.* **2012**, *46*, 1349–1363. [CrossRef] [PubMed]
5. Graham, J.L.; Dubrovsky, N.M.; Eberts, S.M. Cyanobacterial Harmful Algal Blooms and US Geological Survey Science Capabilities. U.S. Geological Survey Report. 2016. Available online: https://pubs.usgs.gov/publication/ofr20161174 (accessed on 23 October 2024).
6. Weirich, C.A.; Miller, T.R. Freshwater harmful algal blooms: Toxins and children's health. *Curr. Probl. Pediatr. Adolesc. Health Care* **2014**, *44*, 2–24. [CrossRef]
7. Yeşilköy, S.; Demir, I. Crop yield prediction based on reanalysis and crop phenology data in the agroclimatic zones. *Theor. Appl. Climatol.* **2024**, *155*, 1–14. [CrossRef]
8. Wells, M.L.; Karlson, B.; Wulff, A.; Kudela, R.; Trick, C.; Asnaghi, V.; Berdalet, E.; Cochlan, W.; Davidson, K.; De Rijcke, M.; et al. Future HAB science: Directions and challenges in a changing climate. *Harmful Algae* **2020**, *91*, 101632. [CrossRef]
9. Tanir, T.; Yildirim, E.; Ferreira, C.M.; Demir, I. Social vulnerability and climate risk assessment for agricultural communities in the United States. *Sci. Total Environ.* **2024**, *908*, 168346. [CrossRef]
10. Greene, S.B.D.; LeFevre, G.H.; Markfort, C.D. Improving the spatial and temporal monitoring of cyanotoxins in Iowa lakes using a multiscale and multi-modal monitoring approach. *Sci. Total Environ.* **2021**, *760*, 143327. [CrossRef]
11. Paerl, H.W.; Gardner, W.S.; Havens, K.E.; Joyner, A.R.; McCarthy, M.J.; Newell, S.E.; Boqiang Qin, B.; Scott, J.T. Mitigating cyanobacterial harmful algal blooms in aquatic ecosystems impacted by climate change and anthropogenic nutrients. *Harmful Algae* **2016**, *54*, 213–222. [CrossRef]
12. Ratté-Fortin, C.; Plante, J.F.; Rousseau, A.N.; Chokmani, K. Parametric versus nonparametric machine learning modelling for conditional density estimation of natural events: Application to harmful algal blooms. *Ecol. Modell.* **2023**, *482*, 110415. [CrossRef]
13. Demiray, B.Z.; Mermer, O.; Baydaroğlu, Ö.; Demir, I. Predicting harmful algal blooms using explainable deep learning models: A comparative study. *Water* **2025**, *17*, 676. [CrossRef]
14. Magnuson, J.J.; Webster, K.E.; Assel, R.A.; Bowser, C.J.; Dillon, P.J.; Eaton, J.G.; Evans, H.E.; Fee, E.J.; Hall, R.I.; Mortsch, L.R.; et al. Potential effects of climate changes on aquatic systems: Laurentian Great Lakes and Precambrian Shield Region. *Hydrol. Process.* **1997**, *11*, 825–871. [CrossRef]
15. Tewari, M.; Kishtawal, C.M.; Moriarty, V.W.; Ray, P.; Singh, T.; Zhang, L.; Treinish, L.; Tewari, K. Improved seasonal prediction of harmful algal blooms in Lake Erie using large-scale climate indices. *Commun. Earth Environ.* **2022**, *3*, 195. [CrossRef]
16. Sterner, R.W.; Keeler, B.; Polasky, S.; Poudel, R.; Rhude, K.; Rogers, M. Ecosystem services of Earth's largest freshwater lakes. *Ecosyst. Serv.* **2020**, *41*, 101046. [CrossRef]
17. Boegehold, A.G.; Burtner, A.M.; Camilleri, A.C.; Carter, G.; DenUyl, P.; Fanslow, D.; Semenyuk, D.F.; Godwin, C.M.; Gossiaux, D.; Johengen, T.H.; et al. Routine monitoring of western Lake Erie to track water quality changes associated with cyanobacterial harmful algal blooms. *Earth Syst. Sci. Data Discuss.* **2023**, *15*, 3853–3868. [CrossRef]
18. Stumpf, R.P.; Johnson, L.T.; Wynne, T.T.; Baker, D.B. Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie. *J. Great Lakes Res.* **2016**, *42*, 1174–1183. [CrossRef]
19. Buratti, F.M.; Manganelli, M.; Vichi, S.; Stefanelli, M.; Scardala, S.; Testai, E.; Funari, E. Cyanotoxins: Producing organisms, occurrence, toxicity, mechanism of action and human health toxicological risk evaluation. *Arch. Toxicol.* **2017**, *91*, 1049–1130. [CrossRef] [PubMed]
20. Carmichael, W.W.; Boyer, G.L. Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes. *Harmful Algae* **2016**, *54*, 194–212. [CrossRef]
21. Kouakou, C.R.; Poder, T.G. Economic impact of harmful algal blooms on human health: A systematic review. *J. Water Health* **2019**, *17*, 499–516. [CrossRef]
22. Wells, M.L.; Trainer, V.L.; Smayda, T.J.; Karlson, B.S.; Trick, C.G.; Kudela, R.M.; Akira Ishikawa, A.; Bernard, S.; Wulff, A.; Anderson, D.M.; et al. Harmful algal blooms and climate change: Learning from the past and present to forecast the future. *Harmful Algae* **2015**, *49*, 68–93. [CrossRef]
23. Glibert, P.M. Harmful algae at the complex nexus of eutrophication and climate change. *Harmful Algae* **2020**, *91*, 101583. [CrossRef]
24. Zhou, Z.X.; Yu, R.C.; Zhou, M.J. Evolution of harmful algal blooms in the East China Sea under eutrophication and warming scenarios. *Water Res.* **2022**, *221*, 118807. [CrossRef] [PubMed]
25. Su, Y.; Hu, M.; Wang, Y.; Zhang, H.; He, C.; Wang, Y.; Wang, D.; Wu, X.; Zhuang, Y.; Hong, S.; et al. Identifying key drivers of harmful algal blooms in a tributary of the Three Gorges Reservoir between different seasons: Causality based on data-driven methods. *Environ. Pollut.* **2022**, *297*, 118759. [CrossRef] [PubMed]
26. Maze, G.; Olascoaga, M.J.; Brand, L. Historical analysis of environmental conditions during Florida Red Tide. *Harmful Algae* **2015**, *50*, 1–7. [CrossRef]
27. Paerl, H.W.; Hall, N.S.; Calandrino, E.S. Controlling harmful cyanobacterial blooms in a world experiencing anthropogenic and climatic-induced change. *Sci. Total Environ.* **2011**, *409*, 1739–1745. [CrossRef]

28. Hushchyna, K.; Sabir, Q.U.A.; Mclellan, K.; Nguyen-Quang, T. Multicollinearity and multi-regression analysis for main drivers of cyanobacterial harmful algal bloom (CHAB) in the Lake Torment, Nova Scotia, Canada. *Environ. Model. Assess.* **2023**, *28*, 1011–1022. [CrossRef]
29. Katin, A.; Del Giudice, D.; Hall, N.S.; Paerl, H.W.; Obenour, D.R. Simulating algal dynamics within a Bayesian framework to evaluate controls on estuary productivity. *Ecol. Modell.* **2021**, *447*, 109497. [CrossRef]
30. Giere, J.; Riley, D.; Nowling, R.J.; McComack, J.; Sander, H. An investigation on machine-learning models for the prediction of cyanobacteria growth. *Fundam. Appl. Limnol.* **2020**, *194*, 85–94. [CrossRef]
31. Greer, B.; McNamee, S.E.; Boots, B.; Cimarelli, L.; Guillebault, D.; Helmi, K.; Marcheggiani, S.; Panaiotov, S.; Breitenbach, U.; Akçaalan, R.; et al. A validated UPLC–MS/MS method for the surveillance of ten aquatic biotoxins in European brackish and freshwater systems. *Harmful Algae* **2016**, *55*, 31–40. [CrossRef]
32. Li, Z.; Xiang, Z.; Demiray, B.Z.; Sit, M.; Demir, I. MA-SARNet: A one-shot nowcasting framework for SAR image prediction with physical driving forces. *ISPRS J. Photogramm. Remote Sens.* **2023**, *205*, 176–190. [CrossRef]
33. Pamula, A.S.; Gholizadeh, H.; Krzmarzick, M.J.; Mausbach, W.E.; Lampert, D.J. A remote sensing tool for near real-time monitoring of harmful algal blooms and turbidity in reservoirs. *J. Am. Water Resour. Assoc. (JAWRA)* **2023**, *59*, 929–949. [CrossRef]
34. Cheng, K.H.; Chan, S.N.; Lee, J.H. Remote sensing of coastal algal blooms using unmanned aerial vehicles (UAVs). *Mar. Pollut. Bull.* **2020**, *152*, 110889. [CrossRef] [PubMed]
35. Kislik, C.; Dronova, I.; Grantham, T.E.; Kelly, M. Mapping algal bloom dynamics in small reservoirs using Sentinel-2 imagery in Google Earth Engine. *Ecol. Indic.* **2022**, *140*, 109041. [CrossRef]
36. Rolim, S.B.A.; Veettil, B.K.; Vieiro, A.P.; Kessler, A.B.; Gonzatti, C. Remote sensing for mapping algal blooms in freshwater lakes: A review. *Environ. Sci. Pollut. Res.* **2023**, *30*, 19602–19616. [CrossRef]
37. Hartshorn, N.; Marimon, Z.; Xuan, Z.; Cormier, J.; Chang, N.B.; Wanielista, M. Complex interactions among nutrients, chlorophyll-a, and microcystins in three stormwater wet detention basins with floating treatment wetlands. *Chemosphere* **2016**, *144*, 408–419. [CrossRef] [PubMed]
38. Hollister, J.W.; Kreakie, B.J. Associations between chlorophyll-a and various microcystin health advisory concentrations. *F1000Research* **2016**, *5*, 151.
39. Zheng, L.; Wang, H.; Liu, C.; Zhang, S.; Ding, A.; Xie, E.; Li, J.; Wang, S. Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models. *J. Environ. Manag.* **2021**, *295*, 113060. [CrossRef]
40. Bui, H.H.; Ha, N.H.; Nguyen, T.N.D.; Nguyen, A.T.; Pham, T.T.H.; Kandasamy, J.; Nguyen, T.V. Integration of SWAT and QUAL2K for water quality modeling in a data scarce basin of Cau River basin in Vietnam. *Ecohydrol. Hydrobiol.* **2019**, *19*, 210–223. [CrossRef]
41. Wool, T.; Ambrose, R.B., Jr.; Martin, J.L.; Comer, A. WASP 8: The next generation in the 50-year evolution of USEPA's water quality model. *Water* **2020**, *12*, 1398. [CrossRef]
42. Shin, C.M.; Kim, D.; Song, Y. Analysis of hydraulic characteristics of Yeongsan River and estuary using EFDC model. *J. Korean Soc. Water Environ.* **2019**, *35*, 580–588.
43. Verhamme, E.M.; Redder, T.M.; Schlea, D.A.; Grush, J.; Bratton, J.F.; DePinto, J.V. Development of the Western Lake Erie Ecosystem Model (WLEEM): Application to connect phosphorus loads to cyanobacteria biomass. *J. Great Lakes Res.* **2016**, *42*, 1193–1205. [CrossRef]
44. Wynne, T.T.; Stumpf, R.P.; Tomlinson, M.C.; Fahnenstiel, G.L.; Dyble, J.; Schwab, D.J.; Joshi, S.J. Evolution of a cyanobacterial bloom forecast system in western Lake Erie: Development and initial evaluation. *J. Great Lakes Res.* **2013**, *39*, 90–99. [CrossRef]
45. Baek, S.S.; Kwon, Y.S.; Pyo, J.; Choi, J.; Kim, Y.O.; Cho, K.H. Identification of influencing factors of *A. catenella* bloom using machine learning and numerical simulation. *Harmful Algae* **2021**, *103*, 102007. [CrossRef]
46. Liu, S.T.; Zhang, L. Surface Chaotic Theory and the Growth of Harmful Algal Bloom. In *Surface Chaos and Its Applications*; Springer: Singapore, 2022; pp. 299–320.
47. Baydaroğlu, Ö.; Yeşilköy, S.; Dave, A.; Linderman, M.; Demir, I. Modeling of Harmful Algal Bloom Dynamics and the Model-Based Interactive Framework for Inland Waters. *EarthArXiv* **2024**, 7075. [CrossRef]
48. Franks, P.J. Recent advances in modelling of harmful algal blooms. In *Global Ecology and Oceanography of Harmful Algal Blooms*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 359–377.
49. Janssen, A.B.; Janse, J.H.; Beusen, A.H.; Chang, M.; Harrison, J.A.; Huttunen, I.; Kong, X.; Rost, J.; Teurlincx, S.; Troost, T.A.; et al. How to model algal blooms in any lake on earth. *Curr. Opin. Environ. Sustain.* **2019**, *36*, 1–10. [CrossRef]
50. Tounsi, A.; Abdelkader, M.; Temimi, M. Assessing the simulation of streamflow with the LSTM model across the continental United States using the MOPEX dataset. *Neural Comput. Appl.* **2023**, *35*, 22469–22486. [CrossRef]
51. Wang, P.; Yao, J.; Wang, G.; Hao, F.; Shrestha, S.; Xue, B.; Xie, G.; Peng, Y. Exploring the application of artificial intelligence technology for identification of water pollution characteristics and tracing the source of water quality pollutants. *Sci. Total Environ.* **2019**, *693*, 133440. [CrossRef]

52. Brehob, M.M.; Pennino, M.J.; Handler, A.M.; Compton, J.E.; Lee, S.S.; Sabo, R.D. Estimates of lake nitrogen, phosphorus, and chlorophyll-a concentrations to characterize harmful algal bloom risk across the United States. *Earth's Future* **2024**, *12*, e2024EF004493. [CrossRef]
53. Yan, Z.; Kamanmalek, S.; Alamdari, N. Predicting coastal harmful algal blooms using integrated data-driven analysis of environmental factors. *Sci. Total Environ.* **2024**, *912*, 169253. [CrossRef]
54. Demir, I.; Beck, M.B. GWIS: A prototype information system for Georgia watersheds. In *Proceedings of the Georgia Water Resources Conference: Regional Water Management Opportunities*; Warnell School of Forestry and Natural Resources, The University of Georgia: Athens, GA, USA, 2009.
55. Demir, I.; Jiang, F.; Walker, R.V.; Parker, A.K.; Beck, M.B. Information systems and social legitimacy: Scientific visualization of water quality. In Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, USA, 11–14 October 2009; IEEE Press: San Antonio, TX, USA, 2009; pp. 1067–1072.
56. Gogtay, N.J.; Thatte, U.M. Principles of correlation analysis. *J. Assoc. Physicians India* **2017**, *65*, 78–81.
57. Wang, H.; Zentar, R.; Wang, D. Predicting the compaction parameters of solidified dredged fine sediments with statistical approach. *Mar. Georesour. Geotechnol.* **2023**, *41*, 195–210. [CrossRef]
58. Tian, D.; Xie, G.; Tian, J.; Tseng, K.H.; Shum, C.K.; Lee, J.; Liang, S. Spatiotemporal variability and environmental factors of harmful algal blooms (HABs) over western Lake Erie. *PLoS ONE* **2017**, *12*, e0179622. [CrossRef]
59. Zhou, Z.X.; Yu, R.C.; Zhou, M.J. Resolving the complex relationship between harmful algal blooms and environmental factors in the coastal waters adjacent to the Changjiang River estuary. *Harmful Algae* **2017**, *62*, 60–72. [CrossRef]
60. Kelley, K.; Bolin, J.H. Multiple regression. In *Handbook of Quantitative Methods for Educational Research*; Brill: Leiden, The Netherlands, 2013; pp. 69–101.
61. Hair, J.F.; Anderson, R.E.; Babin, B.J.; Black, W.C. *Multivariate Data Analysis: A Global Perspective*, 7th ed.; Pearson Education: Upper Saddle River, NJ, USA, 2010.
62. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; García Marquéz, J.R.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
63. Araujo, M.B.; Rahbek, C. How does climate change affect biodiversity? *Science* **2006**, *313*, 1396–1397. [CrossRef]
64. Wheeler, D.C. Geographically weighted regression. In *Handbook of Regional Science*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1895–1921.
65. Helsel, D.R.; Hirsch, R.M. *Statistical Methods in Water Resources*; U.S. Geological Survey, Techniques of Water-Resources Investigations Book 4; 2002. Available online: https://pubs.usgs.gov/publication/tm4A3 (accessed on 17 April 2025).
66. Zuur, A.F.; Ieno, E.N.; Walker, N.J.; Saveliev, A.A.; Smith, G.M. *Mixed Effects Models and Extensions in Ecology with R*; Springer: Berlin/Heidelberg, Germany, 2010.
67. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013.
68. Guisan, A.; Zimmermann, N.E. Predictive habitat distribution models in ecology. *Ecol. Model.* **2000**, *135*, 147–186. [CrossRef]
69. Chorus, I.; Welker, M. (Eds.) *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2021.
70. Igwaran, A.; Kayode, A.J.; Moloantoa, K.M.; Khetsha, Z.P.; Unuofin, J.-L. Cyanobacteria Harmful Algae Blooms: Causes, Impacts, and Risk Management. *Water Air Soil Pollut.* **2024**, *235*, 71. [CrossRef]
71. Havens, K.E. Phosphorus–algal bloom relationships in large lakes of south Florida: Implications for establishing nutrient criteria. *Lake Reserv. Manag.* **2003**, *19*, 222–228. [CrossRef]
72. Zhang, X.; Li, Y.; Zhao, J.; Wang, Y.; Liu, H.; Liu, Q. Temporal dynamics of the chlorophyll a–total phosphorus relationship and algal production efficiency: Drivers and management implications. *Ecol. Indic.* **2024**, *158*, 111339. [CrossRef]
73. Wurtsbaugh, W.A.; Paerl, H.W.; Dodds, W.K. Nutrients, eutrophication and harmful algal blooms along the freshwater to marine continuum. *Wiley Interdiscip. Rev. Water* **2019**, *6*, e1373. [CrossRef]
74. Tao, Y.; Ren, J.; Zhu, H.; Li, J.; Cui, H. Exploring Spatiotemporal Patterns of Algal Cell Density in Lake Dianchi with Explainable Machine Learning. *Environ. Pollut.* **2024**, *356*, 124395. [CrossRef] [PubMed]
75. Lin, S.; Pierson, D.C.; Mesman, J.P. Prediction of algal blooms via data-driven machine learning models: An evaluation using data from a well-monitored mesotrophic lake. *Geosci. Model Dev.* **2023**, *16*, 35–46. [CrossRef]
76. Ai, H.; Zhang, K.; Sun, J.; Zhang, H. Short-term Lake Erie algal bloom prediction by classification and regression models. *Water Res.* **2023**, *232*, 119710. [CrossRef]
77. Izadi, M.; Sultan, M.; Kadiri, R.E.; Ghannadi, A.; Abdelmohsen, K. A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom. *Remote Sens.* **2021**, *13*, 3863. [CrossRef]
78. Qian, J.; Pu, N.; Qian, L.; Xue, X.; Bi, Y.; Norra, S. Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning. *Water Biol. Secur.* **2023**, *2*, 100184. [CrossRef]

79. Gao, L.; Li, X.; Kong, F.; Yu, R.; Guo, Y.; Ren, Y. AlgaeNet: A deep-learning framework to detect floating green algae from optical and SAR imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2782–2796. [CrossRef]
80. Parra, L.; Ahmad, A.; Sendra, S.; Lloret, J.; Lorenz, P. Combination of machine learning and RGB sensors to quantify and classify water turbidity. *Chemosensors* **2024**, *12*, 34. [CrossRef]

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.