Environmental Science & Technology

pubs.acs.org/est

Article

# One-Week-Ahead Prediction of Cyanobacterial Harmful Algal Blooms in Iowa Lakes

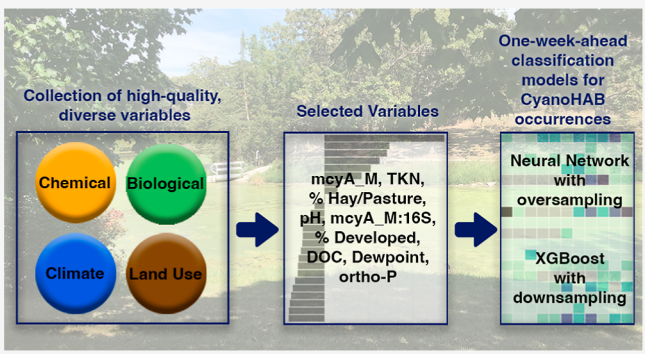
Paul Villanueva, Jihoon Yang, Lorien Radmer, Xuewei Liang, Tania Leung, Kaoru Ikuma, Elizabeth D. Swanner, Adina Howe, and Jaejin Lee*

Cite This: https://doi.org/10.1021/acs.est.3c07764

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Cyanobacterial harmful algal blooms (CyanoHABs) pose serious risks to inland water resources. Despite advancements in our understanding of associated environmental factors and modeling efforts, predicting CyanoHABs remains challenging. Leveraging an integrated water quality data collection effort in Iowa lakes, this study aimed to identify factors associated with hazardous microcystin levels and develop one-week-ahead predictive classification models. Using water samples from 38 Iowa lakes collected between 2018 and 2021, feature selection was conducted considering both linear and nonlinear properties. Subsequently, we developed three model types (Neural Network, XGBoost, and Logistic Regression) with different sampling strategies using the nine selected variables (mcyA_M, TKN, % hay/pasture, pH, mcyA_M:16S, % developed, DOC, dewpoint temperature, and *ortho*-P). Evaluation metrics demonstrated the strong performance of the Neural Network with oversampling (ROC-AUC 0.940, accuracy 0.861, sensitivity 0.857, specificity 0.857, LR+ 5.993, and 1/LR− 5.993), as well as the XGBoost with downsampling (ROC-AUC 0.944, accuracy 0.831, sensitivity 0.928, specificity 0.833, LR+ 5.557, and 1/LR− 11.569). This study exhibited the intricacies of modeling with limited data and class imbalances, underscoring the importance of continuous monitoring and data collection to improve predictive accuracy. Also, the methodologies employed can serve as meaningful references for researchers tackling similar challenges in diverse environments.




**KEYWORDS:** *cyanobacterial harmful algal blooms, microcystin concentration, predictive modeling, freshwater lakes, environmental monitoring, classification models, class imbalance, neural network, XGBoost, logistic regression*

## 1. INTRODUCTION

The growth of microscopic phytoplankton, including Cyanobacteria, on the surface of a water body is natural to some extent.[1] However, under certain environmental, chemical, and biological conditions, their excessive growth can result in harmful algal blooms, which can be highly disruptive to the surrounding environment by altering the physicochemical properties of water bodies, leading to a decline in biodiversity due to oxygen depletion and lack of sunlight.[2−4] Harmful algal blooms are often accompanied by heavy growth of Cyanobacteria (CyanoHABs), which can produce cyanotoxins, such as microcystins and cylindrospermopsin.[4,5] In humans, cyanotoxins can cause a variety of symptoms, including gastrointestinal distress, rashes, liver and kidney toxicity, joint pain, and in extreme cases, neurological damage or paralysis from ingestion or skin contact.[6] Similar exposure to cyanotoxins has also been shown to cause disease or death in wildlife and domestic animals.[4,5] Hence, many national and local governments worldwide have been monitoring inland water resources for cyanotoxins regularly, and the local authorities close the site to the public or recommend avoiding contact when the concentration of cyanotoxins exceeds a certain threshold.[7−9] Subsequently, the closure of beaches or waterways in response to elevated toxin levels may lead to significant economic damage, resulting in lost revenue.[4]

Previous research on CyanoHABs suggest that while blooms have been a long-standing issue,[10] the increased frequency of occurrences is associated with climate change and/or the increasing use of fertilizers in agriculture.[11−13] Historical records show that CyanoHABs have also been a problem in the Midwestern United States for nearly a century. One of the first reports of HABs in the Midwestern states dates from the 1930s, and another report was published in 1951 based on observations of "toxic algae" in Iowa and Minnesota lakes.[10] It has also been recognized that local and global occurrences of

Received: September 22, 2023
Revised: November 1, 2023
Accepted: November 9, 2023

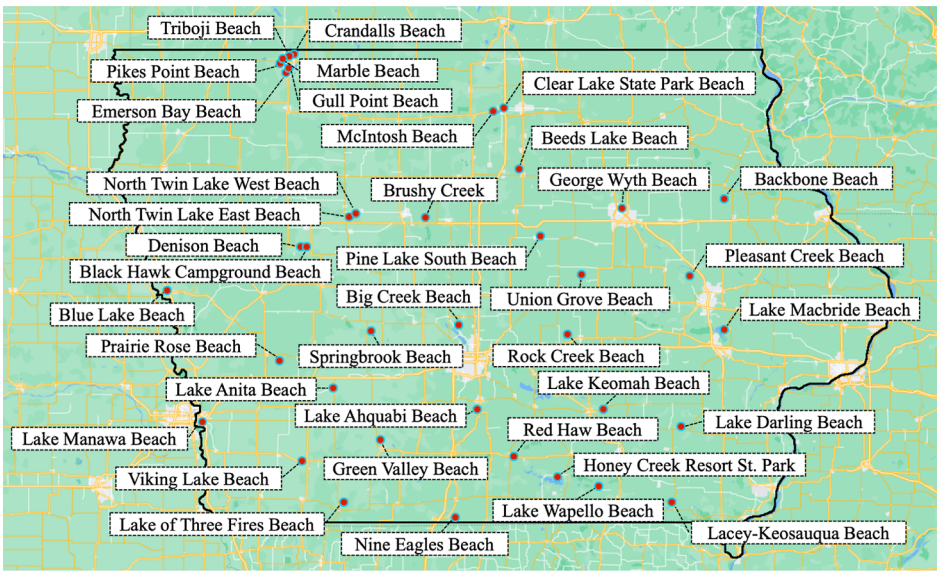



**Figure 1.** Site names and locations of 38 lake water sampling points. Lake water samples were collected in collaboration with the Iowa DNR Beach Monitoring Program from 2018 through 2021.

CyanoHABs are interconnected.[13−15] Water quality challenges in the Midwestern states are not limited to local impact, and connections between the waterways of the Midwestern states and the Mississippi River stream network move nutrients to the Gulf of Mexico.[16,17] The severity and frequency of harmful algal blooms in the Gulf of Mexico have increased since 1985, and total nitrogen loading to the Mississippi River has tripled in about 40 years since the 1950s.[15,18] Despite various efforts in upstream states to reduce nitrogen loading into the Mississippi River and extensive research conducted to understand and mitigate the CyanoHAB problem, negative impacts on the ecosystem and water quality still persist in local lakes and the Gulf of Mexico.[19]

For the last few decades, our understanding of CyanoHAB occurrences has been expanded through numerous studies investigating the relationship with individual factors. For instance, nitrogen and phosphorus levels have been identified as significant factors, as those stimulate cyanobacterial growth.[12,20,21] Phosphorus availability is a known limiting factor for algal growth and has been shown to correlate with harmful algal bloom occurrence.[1,22] Several other factors, such as higher water temperature,[23,24] land use,[25] precipitation,[12,26] wind speeds,[27] and light intensity,[28,29] have also been linked with the growth of toxin-producing Cyanobacteria. However, the correlations with individual factors have not been highly effective in forecasting CyanoHAB occurrences because the activities of toxin-producing Cyanobacteria are controlled by complex interactions of chemical, biological, and climate factors rather than a few individual factors.

Beyond exploring the relationships between CyanoHABs and single factors, plenty of previous studies have applied statistical and machine learning techniques to construct data-driven predictive models for CyanoHABs.[1] Notably, one of the traits commonly found in many previous studies is the lack of clear definitions for the thresholds to determine the occurrence of CyanoHABs in samples (Table S1). The general definition of CyanoHABs is a large growth of microscopic phytoplankton, including Cyanobacteria, on the water surface. However, establishing a standardized definition for CyanoHAB occurrence is crucial for practical predictive modeling because the administrative procedures of local authorities are determined based on the cyanotoxin thresholds recommended by environmental agencies or international health organizations. Therefore, defining CyanoHAB occurrence based on cyanotoxin concentration thresholds advised by authoritative agencies and developing predictive models using cyanotoxin concentration as the target parameter can be more informative for decision-making. Previous studies on forecasting CyanoHABs can be categorized into five groups based on their target parameters: (i) chlorophyll-*a* concentration,[30−35] (ii) cyanobacterial abundance/biomass,[36−40] (iii) phytoplankton biomass,[41−43] and (iv) microcystin concentration,[44−47] and (v) other related parameters[48−52] (Table S1). Most of these studies used chlorophyll-*a* concentration, cyanobacterial abundance/biomass, or phytoplankton biomass as the target parameter to develop predictive models, considering these parameters are proxies for cyanobacterial harmful algal blooms. In some cases, cyanotoxin-producing species like *Microcystis* and *Dolichospermum* were used as the target parameter rather than all Cyanobacteria.[48−52] These predicted targets provide comprehensive information about algae growth in general but may have limitations in directly correlating to CyanoHAB occurrence when the occurrence is defined based on cyanotoxin concentrations. For instance, Kasinak et al. demonstrated a poor correlation between chlorophyll-*a* concentrations and cyanobacterial cell density, which is the bacterial group producing cyanotoxins.[53] A handful of previous studies employed microcystin concentration as the target parameter.[44−47] In a study aiming to issue an early warning if the predicted microcystin concentration exceeds 1 $\mu$g/L, regression models were developed to predict microcystin concentration 10, 20, and 30 days in advance.[44] However, these models did not incorporate diverse parameters as input variables, particularly significant factors such as nitrogen and phosphorus. Similarly, other three previous studies that used microcystin concentration as the target parameter also did not include biological parameters as the input variables for their models.[45−47]

To address these limitations, the current study embarks on the development of a predictive model aimed at enhancing


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

decision-making capabilities concerning CyanoHAB occurrences. This involves utilizing microcystin concentration as the target variable and incorporating various molecular biological, chemical, climate, and land use parameters as input variables. This research was underpinned by collaborative efforts with the Iowa Department of Natural Resources (DNR) between 2018 and 2021, enabling us to generate a comprehensive data set derived from 1591 samples collected across 38 lakes in Iowa. To identify "CyanoHABs" samples, we applied the EPA-recommended microcystin concentration threshold for recreational use, defined as exceeding 8 $\mu$g/L. Since the number of samples surpassing this 8 $\mu$g/L microcystin concentration threshold is considerably smaller compared to the number of "non-CyanoHAB" samples, our approach leaned toward constructing a classification model instead of regression models. In practical terms, this classification model can serve as a diagnostic or screening tool for local authorities, assisting in acknowledging the possibility of microcystin concentrations exceeding the threshold at a site in advance. In essence, our research aims to accomplish two central objectives: (i) the identification of pivotal indicators with a strong predictive capacity for CyanoHABs and (ii) the development of a classification model for forecasting CyanoHAB events with a one-week lead time.

## 2. MATERIALS AND METHODS

**2.1. Sample Collection and Chemical Measurements.** Lake water samples were collected in collaboration with the Iowa DNR Beach Monitoring Program from 38 lake beaches in Iowa, USA (Figure 1). The Iowa DNR has been monitoring various water quality parameters, such as microcystin concentration, dissolved oxygen, and pH, at each lake beach site since 2006. Following the protocol of the Iowa DNR, sampling was conducted on a weekly basis during the summer recreation season from Memorial Day (late May) to Labor Day (early September). The samples used in this study were collected from all 38 lakes between 2018 and 2020. In 2021, as the duration of the EPA project supporting the current study was extended for 1 year due to COVID-19, we decided to collect additional samples to provide more data for model training. We selected one of the four sampling routes, which included lakes with multiple CyanoHAB occurrences in the previous three years, to collect additional samples within the budget. Consequently, 1591 lake water samples were used for further analysis and model training. Detailed information about sites, sampling procedures, and measurements is available on the Beach Monitoring (AQuIA) Web site, operated by the Iowa DNR.[54] Microcystin concentrations for 38 lakes since 2006 were obtained from the AQuIA Web site and used to investigate the trends and occurrences of CyanoHABs over time.

Upon collection, the lake water samples received from the Iowa DNR were immediately stored at 4 °C and analyzed for chemical parameters within 3 days. Prior to collection, pH was determined with a WTW Multi 340i meter and Sentix pH electrode (Weilheim, Germany). According to the EPA 415.3 method, dissolved organic carbon (DOC) was analyzed using a Shimadzu TOC analyzer (Kyoto, Japan) with persulfate digestion. Other chemical parameters were measured using a Seal AQ2 Automated Discrete Analyzer (Seal Analytical, USA). Chloride ($Cl^-$) was measured following the EPA-105-A Rev 5 protocol. Total Kjeldahl nitrogen (TKN) was measured following the EPA-111-A Rev 5 protocol with copper(II) catalyst digestion prior to analysis. Total Kjeldahl phosphorus (TKP) was measured following the EPA-135-A Rev 5 protocol, also using a copper catalyst. Orthophospate (*ortho*-P) was measured following the EPA-118-A Rev 5 protocol with ascorbic acid reduction. The microcystin level in the water samples was determined by Iowa DNR using an enzyme-linked immunosorbent assay (ELISA) test.

**2.2. DNA Extraction and Quantitative Real-Time PCR.** As soon as the samples arrived at the laboratory, 250 mL of each lake water sample was filtered using 0.22 mm PES Membrane filters (Millipore, USA) and stored in a −80 °C freezer until use. From the filters, genomic DNA was extracted using the MagAttract PowerWater DNA/RNA Kit (Qiagen, Germany) and an automated liquid handling system (epMotion 5075, Eppendorf, Germany). To quantify microcystin-producing (*mcy*A) genes in lake water samples, three primer sets (i.e., mcyA_MF/MR for *Microcystis mcy*A genes, mcyA_AF/AR for *Anabaena mcy*A genes, and mcyA_PF/PR for *Planktothrix mcy*A genes) were used,[55] and bacterial 16S rRNA genes were also quantified using 341F/534R primer set.[56] High-throughput qPCR (HT-qPCR) assays were performed on the BioMark HD System (Fluidigm, USA) using Flex Six gene expression IFCs. Each Flex Six IFC contained 48 lake water samples, 24 standards (i.e., 6 serial dilutions for each primer set), and four primer sets in triplicate. According to the manufacturer's protocol, an IFC Controller HX (Fluidigm, USA) was used for priming and loading the IFC. Followed by the default thermal mix and hot start steps, the operating conditions were 40 cycles of (i) 95 °C for 15 s, (ii) 60 °C for 30 s, and (iii) 72 °C for 30 s, followed by the melting step. After the number of target genes per reaction was calculated from the corresponding standard curve, the number of target genes per mL of the sample was determined using the volume of extracted DNA, the volume of DNA in each reaction, and the volume of the filtered sample.[57]

**2.3. Land Use and Climate Data.** Land use information was obtained from the National Land Cover Database (NLCD).[58] The latest version of land cover classifications was released in 2019. The percentage of land dedicated to each land use category within 1 km were determined for each sampling site using the raster, sf, and exactextractr packages in the R programming and added as additional variables for each observation.[59−61] Ordination via nonmetric dimensional scaling was applied to sampling sites based on their land-use profiles to determine if there were any clustering patterns apparent between sites (Figure S1).

Climate data were sourced from Weather Underground, an online network of local weather stations.[62] We collected weekly averages of temperature, humidity, dewpoint temperature, wind speed, gust speed, and precipitation from the weather station nearest to each sampling site. In cases where data could not be obtained from the closest weather station to a sampling point, we utilized readings from the next closest station. Since the locations of these weather stations were not originally intended for collecting climate data from lake beaches, there were variations in the distances between the sampling points and the closest weather stations (Table S2).

**2.4. Data Preparation.** The parameters collected in this study included microcystin concentrations, pH, DOC, $Cl^-$, TKN, TKP, *ortho*-P, *Microcystis mcy*A gene copies (mcyA_M), *Planktothrix mcy*A gene copies (mcyA_P), *Anabaena mcy*A gene copies (mcyA_A), bacterial 16S rRNA gene copies (16S rRNA), precipitation, temperature, dewpoint temperature,


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

wind speed, gust speed, humidity, and percentages of land use within a 1 km radius from each sampling point (for instance,% developed, % hay/pasture, % cultivated crops, and others) (Table S3). Based on the EPA-recommended microcystin concentrations for recreational use (below 8 $\mu$g/L), the observed microcystin concentrations in lake water samples were categorized into two groups: "Hazardous" and "Safe". Among the 1591 samples, 79 samples were labeled as "Hazardous" while 1512 samples were labeled as "Safe". Since the data were collected weekly, each week's input variables were paired with the subsequent week's microcystin safety level (i.e., Hazardous or Safe) to build one-week-ahead predictive models, with the observed one-week-ahead concentration serving as the target prediction for the models (Figure S2). Once the input variables and the target variable (i.e., the following week's microcystin safety level) were linked together, the sequential property of the original data set was no longer considered.[63] Following the rearrangement of the original data set by pairing the input variables of the prior week with the microcystin safety level of the following week, a total of 1473 pairs of input and target variables were used for further analysis. Among these, 70 cases were labeled as "Hazardous", and 1403 were labeled as "Safe". Wilcoxon rank-sum tests were conducted to assess the significance of differences in the mean values of the input variables between the Hazardous and Safe groups. For machine learning, the data were divided into training and testing sets using an 80:20 split stratified based on the following week's microcystin safety level. The division resulted in a training set with 56 Hazardous cases with 1122 Safe cases, while the testing set contained 14 Hazardous and 281 Safe cases. For the training set, two separate procedures (oversampling and downsampling) equalizing the proportion of classes were implemented to address the class imbalance. In the downsampling procedure, the majority (Safe) class was randomly downsampled to match the number of observations of the minority (Hazardous) class. The oversampling procedure generated simulated observations of the Hazardous class by following the SMOTE algorithm until the number of observations in the minority class equaled that of the majority class.[64] The initial training set without adjustments was also retained for subsequent procedures. These three different training sets were used in separate configurations of the model training pipeline and evaluated separately. All the data used in this study and subsequent analyses are publicly archived at https://github.com/pommevilla/one_week_ahead.

**2.5. Feature Selection.** Feature selection was conducted before model training to ensure consistent feature sets and enhance model performance evaluation for predicting Hazardous and Safe classes. After creating 1000 permutations of the data set by sampling from the original data set 1000 times with replacement, a LASSO model and an XGBoost model were trained using glmnet[65] and XGboost[66] packages, respectively. LASSO is a linear model that excels at inducing sparsity by eliminating features, while XGBoost excels at capturing complex relationships through nonlinear modeling. The aim of combining these two techniques was to create a more comprehensive feature selection approach capable of ultimately enhancing the prediction performance. Importance scores were assigned to each feature during the training process for the feature selection, with 1000 scores generated for each feature from both the LASSO and the XGBoost models after training. For each of the 1000 permutations, the feature importance scores of the two models were normalized separately to a mean of 0 and a standard deviation of 1. An overall average importance score was calculated for each feature by calculating the mean LASSO importance and mean XGBoost importance and then taking the mean of the two scores. The overall average importance score, akin to a $Z$-score, serves as an indicator of the predictive power of each feature. A score of 0 denotes average predictive power; scores below 0 suggest below-average predictive power, and scores above 0 indicate above-average predictive power. Therefore, features with an overall average importance score of >0 were chosen for constructing the final models.

**2.6. Model Training and Performance Assessment.** To evaluate their effectiveness in predicting CyanoHAB occurrence in the following week, three types of models, including XGBoost, Neural Network (Feedforward Neural Networks), and Logistic Regression (Elastic Net), were trained based on the selected features using XGboost,[66] Brulee,[67] and glmnet[65] packages, respectively. For each model, hyperparameter tuning was first performed by (i) creating a set of candidate hyperparameter values for each model, (ii) creating a model for each combination of hyperparameter values, (iii) training each model on the training set using 10-fold cross validation, and (iv) recording the average receiver operating characteristic area under the curve (ROC-AUC) (i.e., overall metric indicating how well the model does at discriminating between classes) across all folds on the testing set (Table S4). For each model type, the set of hyperparameters achieving the highest average ROC-AUC was kept and used to compare the different models against each other. Among the 200 candidates for each model type, the candidate model that achieved the highest ROC-AUC on the testing set was selected as the representative for each model type (Table S5). The performance evaluation of each representative model was conducted using the testing set (14 Hazardous and 281 Safe cases) based on various performance metrics, including ROC-AUC, accuracy (i.e., the ratio of correct predictions out of all predictions), sensitivity (i.e., the ratio of correct predictions for the samples belonging to the Hazardous class), specificity (i.e., the ratio of correct predictions for the samples belonging to the Safe class). To intuitively present the overall performance of each model, likelihood ratios (i.e., ratios to compare the probability of an event happening to the probability of it not happening) were used, which are often employed in hypothesis testing and diagnostic testing.[68] Likelihood ratios are represented as the likelihood ratio for a positive test result (LR+) and the likelihood ratio for a negative test result (LR−). A likelihood ratio of 1.0 means no difference in the probability of the particular test results between the positive and negative cases and, for example, LR+ = $x$ means that a sample with >8 $\mu$g/L microcystin concentration is $x$ times more likely to be predicted as "Hazardous" than a sample with <8 $\mu$g/L microcystin concentration. Therefore, models with higher LR + and or higher 1/LR− (i.e., lower LR−) can be expected to exhibit better performance.

## 3. RESULTS

**3.1. Unpredictable Patterns of CyanoHAB Occurrences in Iowa Lakes.** Among the 1591 samples collected between 2018 and 2021, 79 samples exceeded the EPA threshold for recreational use (i.e., 8 $\mu$g/L) and were labeled as Hazardous. Notably, 50 of the Hazardous samples were collected from three locations: Lake Darling Beach (19 occurrences), Green Valley Beach (18 occurrences), and


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

**Table 1. Number of Cyanobacterial Harmful Algal Bloom Occurrences (i.e., Exceeding 8 μg/L Microcystin Concentration Based on the EPA-Recommended Threshold for Recreational Use) by Location and Year**

| location | 2018 | 2019 | 2020 | 2021 | total | location | 2018 | 2019 | 2020 | 2021 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lake Darling Beach | 0 | 9 | 2 | 8 | 19 | Black Hawk Campground Beach | 0 | 0 | | | 0 |
| Green Valley Beach | 3 | 10 | 0 | 5 | 18 | Brushy Creek | 0 | 0 | 0 | | 0 |
| Lake of Three Fires Beach | 3 | 6 | 3 | 1 | 13 | Emerson Bay Beach | 0 | 0 | 0 | | 0 |
| Union Grove Beach | 0 | 2 | 5 | | 7 | Gull Point Beach | 0 | 0 | 0 | | 0 |
| Viking Lake Beach | 7 | 0 | 0 | 0 | 7 | Lacey-Keosauqua Beach | 0 | 0 | 0 | | 0 |
| Lake Keomah Beach | 0 | 3 | 0 | 0 | 3 | Lake Ahquabi Beach | 0 | 0 | 0 | | 0 |
| McIntosh Beach | 1 | 1 | 0 | | 2 | Lake Anita Beach | 0 | 0 | 0 | | 0 |
| Beeds Lake Beach | 0 | 1 | 0 | | 1 | Lake Manawa Beach | 0 | 0 | 0 | | 0 |
| Clear Lake State Park Beach | 0 | 1 | 0 | | 1 | Lake Wapello Beach | 0 | 0 | 0 | 0 | 0 |
| Crandalls Beach | 1 | 0 | 0 | | 1 | Blue Lake Beach | 0 | 0 | 0 | | 0 |
| Denison Beach | 1 | 0 | 0 | | 1 | Pine Lake South Beach | 0 | 0 | 0 | | 0 |
| George Wyth Beach | 0 | 0 | 1 | | 1 | Nine Eagles Beach | 0 | 0 | 0 | | 0 |
| Honey Creek Resort St. Park | 0 | 1 | 0 | | 1 | North Twin Lake West Beach | 0 | 0 | 0 | | 0 |
| Lake Macbride Beach | 1 | 0 | 0 | 0 | 1 | Pikes Point Beach | 0 | 0 | 0 | | 0 |
| Marble Beach | | | 1 | | 1 | Pleasant Creek Beach | 0 | 0 | 0 | | 0 |
| North Twin Lake East Beach | 0 | 1 | 0 | | 1 | Prairie Rose Beach | 0 | 0 | 0 | | 0 |
| Rock Creek Beach | 0 | 0 | 0 | 1 | 1 | Red Haw Beach | 0 | 0 | 0 | | 0 |
| Backbone Beach | 0 | 0 | 0 | | 0 | Springbrook Beach | 0 | 0 | 0 | | 0 |
| Big Creek Beach | 0 | 0 | 0 | | 0 | Triboji Beach | 0 | 0 | 0 | | 0 |

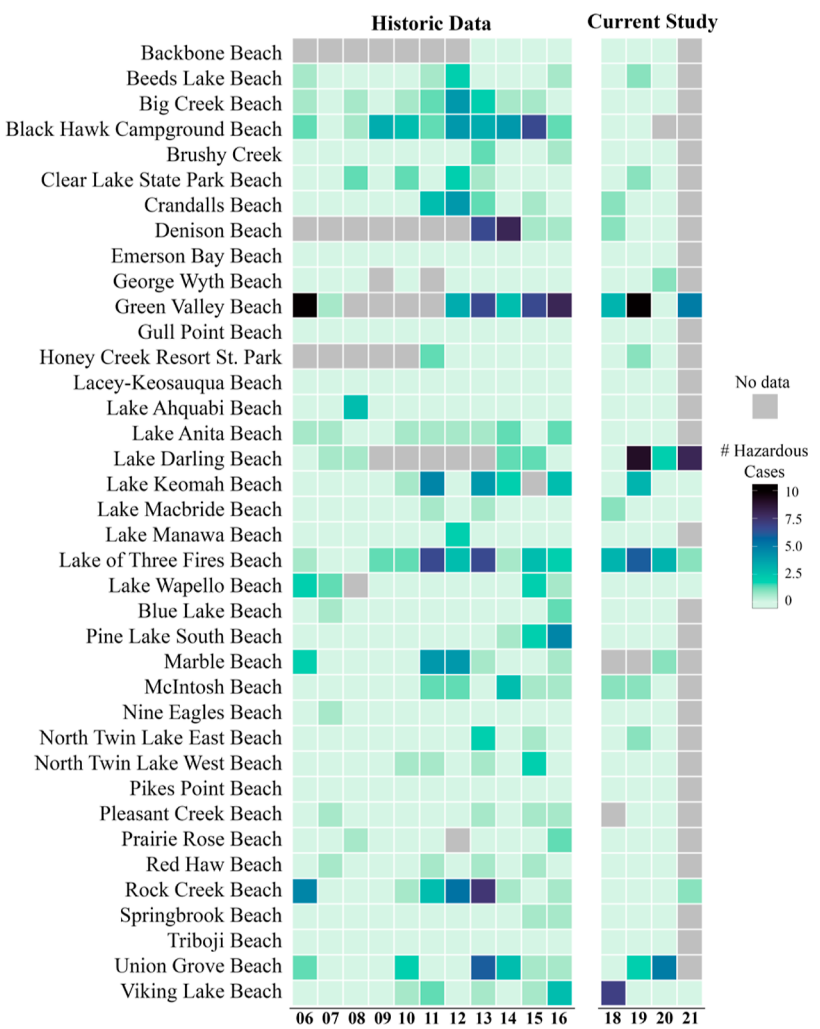



**Figure 2.** Count of CyanoHAB occurrences (i.e., based on the current EPA threshold for recreational water, 8 μg/L of microcystins) at Iowa lakes between Memorial Day and Labor Day for each year since 2006.



https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 2. Summary of Wilcoxon Rank-Sum Tests Conducted on All Available Variables between Hazardous and Safe Samples[a]

| category | variable | hazardous ($n$ = 70) | safe ($n$ = 1403) | $P$-value |
|---|---|---|---|---|
| chemical | **pH** | **9.05** | **8.46** | **<0.001*** ** |
| | **DOC (ppm)** | **8.64** | **6.35** | **<0.001*** ** |
| | **$Cl^-$ (mg/L)** | **10.44** | **14.82** | **<0.001*** ** |
| | **TKN (mg N/L)** | **2.30** | **0.86** | **<0.001 *** ** |
| | TKP (mg P/L) | 0.51 | 0.50 | 0.5 |
| | ***ortho*-P (mg P/L)** | **0.14** | **0.04** | **<0.001*** ** |
| | TP (mg P/L) | 0.64 | 0.54 | 0.076 |
| biological | **Microcystis *mcy*A (copies/mL)** | **$1.30 \times 10^6$** | **$7.38 \times 10^4$** | **<0.001*** ** |
| | Planktothrix *mcy*A (copies/mL) | $6.43 \times 10^3$ | $4.23 \times 10^3$ | 0.7 |
| | Anabaena *mcy*A A (copies/mL) | $0.00 \times 10^0$ | $0.00 \times 10^0$ | >0.9 |
| | 16s rRNA (copies/mL) | $7.02 \times 10^6$ | $6.19 \times 10^6$ | 0.6 |
| | ***mcy*A_M:16s rRNA** | **0.18** | **0.02** | **<0.001*** ** |
| climate | precipitation (mm) | 4.06 | 3.3 | 0.8 |
| | temperature (°C) | 22.7 | 22.4 | 0.6 |
| | dewpoint (°C) | 17.2 | 17.8 | 0.7 |
| | wind speed (m/s) | 1.02 | 1.03 | 0.6 |
| | gust speed (m/s) | 1.88 | 1.97 | 0.8 |
| | humidity (%) | 75 | 76 | >0.9 |
| land use | **% wetlands** | **3.0** | **6.0** | **<0.001*** ** |
| | % forest | 29.0 | 23.0 | 0.003 |
| | **% developed** | **7.0** | **11.0** | **<0.001*** ** |
| | % barren land | 0.1 | 0.1 | 0.231 |
| | % scrub/shrubbery | 0.6 | 0.3 | 0.010 |
| | % cultivated crops | 20.0 | 16.0 | 0.027 |
| | % open water | 18.0 | 30.0 | 0.003 |
| | **% hay/pasture** | **19.0** | **10.0** | **<0.001*** ** |
| | % herbaceous | 3.0 | 3.0 | 0.2 |

[a]The bold rows indicate the variables with a statistically significant difference in means between the two classes ($P$-value <0.001).

Lake of Three Fires Beach (13 occurrences) (Table 1). When also considering the number of Hazardous cases that had occurred in Union Grove Beach (7 occurrences) and Viking Lake Beach (7 occurrences), 81% (64/79) of all CyanoHAB occurrences were observed in five out of 38 lakes. The observations in the study period may suggest that there are certain lakes with repetitive problems; however, the historic data gathered from 2006 does not support this hypothesis (Figure 2).[54] Out of 38 lakes, only five lake beaches have been consistently free of CyanoHAB problems since 2006 and most of those have had CyanoHAB occurrences from time to time without distinctive patterns. For instance, the historic data show that Black Hawk Campground Beach had many Hazardous occurrences between 2008 and 2015 but had no recorded instances of CyanoHAB occurrences during the study period. Viking Lake, which had the highest number of Hazardous samples in 2018, had no such cases in 2019, 2020, and 2021. Conversely, Union Grove had no Hazardous cases in 2018, but had two cases in 2019 and five cases in 2020. Overall, the monitoring data suggest that the patterns of CyanoHAB occurrences are neither predictable nor consistent.

**3.2. Identifying Key Factors for Predicting Cyano-HABs in Iowa Lakes.** Differences between all chemical, biological, climate, and land use measurements in Hazardous and Safe groups were compared and significant differences ($p$ < 0.001) were observed between the Hazardous and Safe groups for 10 of the 27 variables analyzed using Wilcoxon rank-sum tests on the rearranged data set that each week's observations were coupled with the following week's microcystin safety level ($n$ = 1473) (Table 2). These variables included pH, DOC, $Cl^-$, TKN, *ortho*-P, mcyA_M, mcyA_M:16S, percentages of land within a 1 km radius from the sampling point categorized as wetlands (% wetlands), developed area (% developed), and hay/pasture area (% hay/pasture). The gene copies of *mcy*A genes (mcyA_M), which are directly associated with microcystin production,[69] were observed to be approximately 20 times higher than in Hazardous cases compared to safe samples. Similarly, the mean ratio of *Microcystis mcy*A to bacterial 16S rRNA gene copies (mcyA_M:16S) was as high as 0.18 in the Hazardous group while the ratio was 0.02 in the Safe group. Consistent with observations that high nutrients loads are associated with CyanoHABs, the concentrations of nutrients such as TKN and *ortho*-P were significantly higher in the Hazardous group. However, no significant differences were observed in TKP and TP measurements between Safe and Hazardous groups. Other chemical parameters like DOC and pH were also higher in the Hazardous cases, while the mean chloride concentration was higher in the Safe group.

Through the feature selection on the rearranged data set that linked each week's input variables together with the following week's microcystin safety level class ($n$ = 1473), nine factors were identified as having above-average predictive power based on the average normalized feature importance scores obtained (Figure 3). The feature selection analysis suggested that microcystin-producing gene copies of *Microcystis* (mcyA_M) were the most important factor in predicting CyanoHAB occurrences and chose multiple parameters from all available categories, including chemical, biological, climate, and land use variables. These highly predictive factors, including mcyA_M, TKN, % hay/pasture, pH, mcyA_M:16S, % developed, DOC, dewpoint temperature, and *ortho*-P, were selected as input variables to train subsequent models.


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

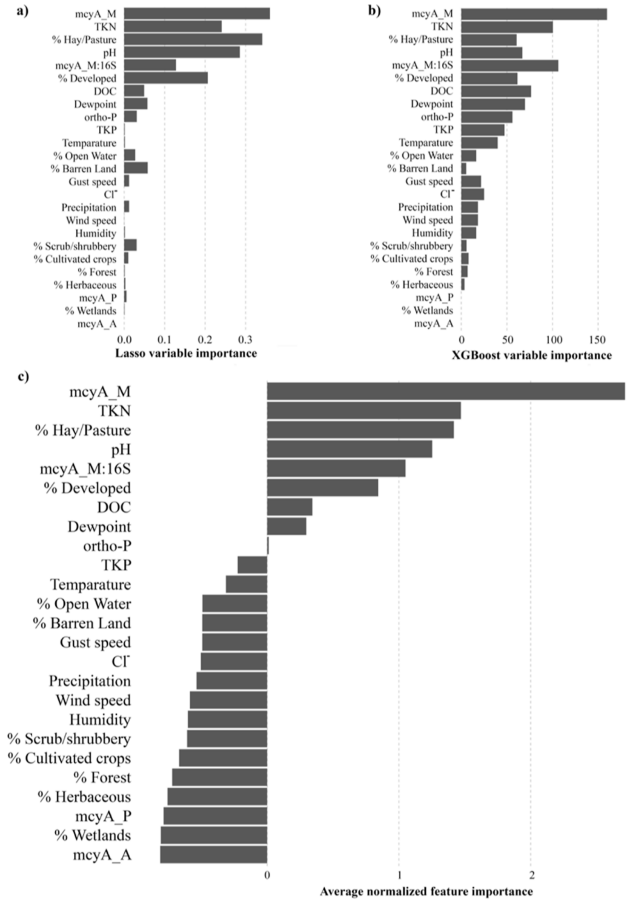



**Figure 3.** Variable importance scores as determined by average over 1000 iterations of (a) LASSO and (b) XGBoost on permuted data sets. The LASSO and XGBoost scores were normalized separately, then averaged together to derive (c) an average normalized importance score. A score of 0 indicates that the feature has average predictive power, a negative score indicates predictive power worse than average, and positive scores indicate better-than-average predictive power.

**3.3. Performance Evaluation of One-Week-Ahead CyanoHABs Prediction Models.** The present study evaluated three different model types (XGBoost, Neural Network, and Logistic Regression) with three different sampling strategies to train the selected variables. The performance of each model on the testing set was compared using diverse metrics, including ROC-AUC, accuracy, sensitivity, specificity, LR+, and 1/LR− (Table 3). The three models for the training set without class imbalance adjustment showed high scores in ROC-AUC, accuracy, and specificity, while the sensitivity scores were low (0.357 for XGBoost, 0.286 for Neural Network, and 0.000 for Logistic Regression). As the low sensitivity scores indicate that the models are not capable of predicting Hazardous cases correctly, the models built on the training set without class imbalance adjustment were excluded from further considerations. The Neural Network model with oversampling most consistently demonstrated high scores across all evaluation metrics (i.e., ROC-AUC 0.940, accuracy 0.861, sensitivity 0.857, specificity 0.857, LR+ 5.993, and 1/LR− 5.993). Likelihood ratios of the Neural Network model with oversampling indicate that a Hazardous case is 5.993 times more likely to be predicted as "Hazardous" than a Safe case, while a Safe case is 5.993 times more likely to be predicted as "Safe" than a Hazardous case. Interestingly, the LR+ of the XGBoost with downsampling (5.557) was slightly lower than the LR+ of the Neural Network model with downsampling (5.993), while it showed a much higher 1/LR− (11.569) value than the Neural Network model with downsampling (5.993). In other words, through the XGBoost model with downsampling, a Hazardous case is 5.557 times more likely to be predicted as "Hazardous" than a Safe case, while a Safe case is 11.569 times more likely to be predicted as "Safe" than a Hazardous case. Other models, such as the XGBoost with oversampling, the Neural Network with downsampling, and Logistic Regression models, showed a bit lower LR + or 1/LR− compared to the XGBoost with downsampling and the Neural Network oversampling. While both the XGBoost with downsampling and the Logistic Regression model with oversampling achieved the highest sensitivity scores, correctly predicting 13 out of 14 Hazardous cases, the Logistic Regression model had more false positives, resulting in a lower LR+. In summary, the evaluation of one-week-ahead CyanoHABs prediction models consistently demonstrated the strong performance of the Neural Network model with oversampling. Notably, XGBoost with downsampling and the Neural Network with oversampling emerged as strong contenders, each excelling in different aspects of predictive power. The XGBoost with downsampling exhibited a slightly lower LR+ but a significantly higher 1/LR−, while

**Table 3. Performance Metrics on the Test Set for the Models Trained**[a]

| model | sampling strategy | ROC-AUC | accuracy | sensitivity | specificity | LR+ | 1/LR− |
|---|---|---|---|---|---|---|---|
| XGBoost | oversampling | **0.940** | **0.902** | 0.786 | **0.900** | **7.860** | 4.206 |
| | downsampling | **0.944** | **0.831** | **0.928** | 0.833 | **5.557** | **11.569** |
| | none | 0.930 | 0.963 | 0.357 | 0.993 | | |
| neural network | oversampling | **0.940** | **0.861** | **0.857** | **0.857** | **5.993** | **5.993** |
| | downsampling | 0.932 | 0.810 | 0.786 | 0.807 | 4.073 | 3.771 |
| | none | 0.925 | 0.966 | 0.286 | 1.000 | | |
| logistic regression | oversampling | 0.932 | 0.813 | **0.928** | 0.811 | 4.910 | **11.264** |
| | downsampling | 0.915 | **0.831** | 0.714 | **0.839** | 4.435 | 2.934 |
| | none | 0.927 | 0.953 | 0.000 | 1.000 | | |

[a]The sampling strategy refers to how class imbalances were handled: oversampling refers to using the SMOTE algorithm to generate observations of the minority class to even out class proportions, downsampling refers to randomly selecting samples in the majority class as many as the number of samples in the minority class, and none refers to the training set without class imbalance adjustment. After excluding the models with low sensitivity scores, which were built on the training set without class imbalance adjustment, the top three scores in each metric are displayed in bold boxes.


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

the Neural Network with oversampling demonstrated a balanced performance across all metrics.

## 4. DISCUSSION

The primary objectives of the current study were to identify predictive factors for CyanoHAB occurrences and to develop classification models capable of one-week-ahead forecasting whether microcystin concentrations would exceed EPA thresholds. Feature selection highlighted the significance of nine key factors, spanning biological (mcyA_M, mcyA_M:16S), chemical (TKN, pH, DOC, *ortho*-P), land use (% hay/pasture, % developed), and climatic (dewpoint temperature) variables. The machine learning model evaluation, based on six performance metrics (ROC-AUC, accuracy, sensitivity, specificity, LR+, and 1/LR−), led to the recommendation of the Neural Network with oversampling and the XGBoost with downsampling. These models offer stable and balanced predictions of CyanoHAB occurrences. Overall, this study underscores the potential of machine learning approaches in predicting CyanoHABs and emphasizes the importance of an integrative approach to understanding the complex interplay of variables influencing CyanoHAB occurrences.

One of the principal findings of this study is the close alignment between the relationships between CyanoHABs and individual factors described in the existing literature and the variables that displayed significant differences between Hazardous and Safe groups in Wilcoxon rank-sum tests (Table 2), which are also well aligned with the input variables selected for model training (Figure 3). For example, samples exceeding the EPA threshold (i.e., 8 $\mu$g/L microcystin concentration) in the following week exhibited significantly higher nitrogen and phosphorus concentrations, corroborating previous studies.[20,21,39] Earlier research have also demonstrated a strong relationship between changes in *Microcystis* activities and DOC concentrations or pH.[70] Notably, *ortho*-P, which can be directly utilized by microscopic phytoplankton,[71] emerged as a more predictive variable than TKP. While only dewpoint temperature was chosen for model training among the climate parameters, despite no significant differences observed between the Hazardous and Safe groups in this regard, there is room for discussion regarding why the feature selection algorithms singled out dewpoint temperature. A possible explanation might be found in a previous study conducted in Cobscook Bay, Maine, USA,[72] which reported a negative correlation between higher dewpoint temperatures and harmful algal blooms. This correlation could be attributed to higher dewpoint temperatures indicating cloudier or stormier weather conditions, characterized by increased atmospheric water vapor, which is less favorable for cyanobacterial growth. Alternatively, the consistency of weather conditions across Iowa or the varying distances from weather stations to the sampling sites (Table S2) could have influenced the statistical significance of the climate parameters.

According to the National Land Cover Database (NLCD), planted/cultivated land can be categorized into two classes: (i) % hay/pasture, which refers to areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, and (ii) % cultivated crops, which represents areas used for the production of annual crops, such as corn, soybeans, vegetables, and more. Since agricultural land is generally considered a major source of nutrient loads into inland water resources, both types of land use showed statistical difference between Hazardous and Safe groups (Table 2), although the significance was weaker for cultivated crop area ($p < 0.05$). However, the overall analysis, including feature selection, suggest that the % hay/pasture area can be a more influential source of nutrient loads into the lakes compared to % Cultivated Crop. This higher influence of % hay/pasture has also been demonstrated in a previous study, which suggested decreasing the percentage of grazing land cover can potentially improve water quality and reduce the occurrence of cyanobacterial blooms in water reservoirs.[73] In the same context, the negative correlations with % developed and % wetland between Hazardous and Safe groups can be explained, as a larger developed and wetland area indicates less agricultural area nearby.

The successful development of accurate data-driven models for natural events in the environment may require four essential elements: (i) the collection of high-quality data sets encompassing diverse and relevant variables, (ii) the collection of sufficient data, (iii) the selection of appropriate machine learning algorithm(s), and (iv) the formulation of a well-defined research question.[74,75] With the assumption that a well-defined research question can only emerge from a proper data set, the question arises: among a high-quality data set containing diverse and relevant input variables, a sufficient amount of data, and appropriate machine learning algorithms, which element holds the utmost importance? A previous study on rainfall-runoff modeling, which assessed the influence of input data, model type, preprocessing, and data length on forecasting accuracy, concluded that the primary element is the input data, followed by data length, preprocessing, and model type.[76] Unfortunately, attempts to construct data-driven models using environmental data sets often get caught in the trap of comparing various model types with an inadequate list of variables and searching for ways to modify the data to achieve better performance. This undesirable initial approach can also lead to the adoption of indirect parameters as target values, frequently resulting in a disconnect from effectively addressing the ultimate research question. A sufficient amount of high-quality, relevant data with appropriate training can enable machine learning models to achieve desired levels of accuracy for desired target variables.[77]

In the case of predicting CyanoHABs, the ideal scenario would involve developing a regression model that forecasts specific concentrations of microcystins or other cyanotoxins with some lead time. Although the current study collected 1591 samples from dozens of lakes across Iowa in collaboration with a state government agency, the data set was not suitable for regression models due to the limited number of samples with microcystin concentrations above 8 $\mu$g/L (Figure S3). Consequently, the target parameter had to be adjusted by categorizing the microcystin concentrations into two classes according to the EPA threshold (i.e., 8 $\mu$g/L). Despite this limitation, our models still serve as a practical solution from the regulatory perspective by informing us whether a site should be closed, even without informing specific microcystin concentration values. Another limitation was the class imbalance in the available data, with the majority of lake water samples classified as Safe. This imbalance skews the model performance as a model predicting all cases to be the majority class would achieve high accuracy and sensitivity but low specificity (i.e., less correctly classifying samples belonging to the minority class). Excluding data from lakes without CyanoHABs occurrence during the study period cannot be


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

considered as an option to improve the class imbalance. Such data from lakes without historical issues are equally important as they can provide information about why certain lakes have remained problem-free. Additionally, as shown in Figure 2, all lakes have the potential for CyanoHABs occurrences if specific conditions are met. Thus, the class imbalance problem needs to be addressed further, for example, by collecting more Hazardous cases through continuous and long-term sampling efforts. Such continuous monitoring efforts will enhance our understanding of the dynamics surrounding CyanoHAB occurrences and facilitate practical mitigation strategies. Additionally, further investigations are also necessary to understand the intricate interplay among relevant variables in both laboratory and field settings. These investigations should involve identifying parameters that serve as root causes or triggers or simply exhibit co-occurrences.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c07764.

> Literature review on previous CyanoHABs prediction efforts; distances between the sampling points and the weather stations; description of parameters collected; parameters tuned during model training; hyperparameters for each model; nonmetric dimensional scaling of sampling sites based on land-use classifications; visual explanation of data rearrangement of the data set to pair input variables from the prior week with the microcystin safety level for the subsequent week; and distribution of microcystin concentrations in 1591 lake water samples (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Jaejin Lee** − *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0002-9793-9473; Phone: +1-515-294-1434; Email: jaylee@iastate.edu

### Authors

**Paul Villanueva** − *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011, United States*

**Jihoon Yang** − *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0002-7018-121X

**Lorien Radmer** − *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011, United States*

**Xuewei Liang** − *Department of Civil, Construction and Environmental Engineering, Iowa State University, Ames, Iowa 50011, United States*

**Tania Leung** − *Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0002-8218-1313

**Kaoru Ikuma** − *Department of Civil, Construction and Environmental Engineering, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0003-3715-7821

**Elizabeth D. Swanner** − *Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa 50011, United States;* orcid.org/0000-0001-9507-0893

**Adina Howe** − *Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c07764


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the US Environmental Protection Agency [EPA-G2017-STAR-A1]. We are also grateful to Daniel Kendall and Amy Buckendahl of the Iowa Department of Natural Resources for providing lake water samples across Iowa.

## ■ REFERENCES


(1) Rousso, B. Z.; Bertone, E.; Stewart, R.; Hamilton, D. P. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Res.* **2020**, *182*, 115959.

(2) Paerl, H. W.; Huisman, J. Blooms like it hot. *Science* **2008**, *320* (5872), 57−58.

(3) Smayda, T. J. What is a bloom? A commentary. *Limnol. Oceanogr.* **1997**, *42*, 1132−1136.

(4) 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.

(5) Carmichael, W. W. The toxins of Cyanobacteria. *Sci. Am.* **1994**, *270* (1), 78−86.

(6) Zurawell, R. W.; Chen, H.; Burke, J. M.; Prepas, E. E. Hepatotoxic cyanobacteria: a review of the biological importance of microcystins in freshwater environments. *J. Toxicol. Environ. Health, Part B* **2005**, *8* (1), 1−37.

(7) World Health Organization. *Toxic Cyanobacteria in Water—Second Edition*, 2021. https://www.who.int/publications/m/item/toxic-cyanobacteria-in-water—second-edition.

(8) United States Environmental Protection Agency. *Final Technical Support Document: Implementing the 2019 National Clean Water Act Section 304(a) Recommended Human Health Recreational Ambient Water Quality Criteria or Wwimming Advisories for Microcystins and Cylindrospermopsin. EPA 823-R-21-002*, 2021. https://www.epa.gov/system/files/documents/2021-08/final-tsd-implement-2019-rwqc.pdf.pdf.

(9) Centers for Disease Control and Prevention. *Facts about Cyanobacterial Blooms for Poison Center Professionals*, 2022. https://www.cdc.gov/habs/materials/factsheet-cyanobacterial-habs.html.

(10) Rose, E. T. Toxic algae in Iowa lakes. *Proc. Iowa Acad. Sci.* **1953**, *60* (1), 738−746.

(11) Weber, S. J.; Mishra, D. R.; Wilde, S. B.; Kramer, E. Risks for cyanobacterial harmful algal blooms due to land management and climate interactions. *Sci. Total Environ.* **2020**, *703*, 134608.

(12) Wells, M. L.; Karlson, B.; Wulff, A.; Kudela, R.; Trick, C.; Asnaghi, V.; Berdalet, E.; Cochlan, W.; Davidson, K.; De Rijcke, M.; Dutkiewicz, S.; Hallegraeff, G.; Flynn, K. J.; Legrand, C.; Paerl, H.; Silke, J.; Suikkanen, S.; Thompson, P.; Trainer, V. L. Future HAB science: Directions and challenges in a changing climate. *Harmful Algae* **2020**, *91*, 101632.

(13) Mrdjen, I.; Fennessy, S.; Schaal, A.; Dennis, R.; Slonczewski, J. L.; Lee, S.; Lee, J. Tile drainage and anthropogenic land use contribute to harmful algal blooms and microbiota shifts in inland water bodies. *Environ. Sci. Technol.* **2018**, *52* (15), 8215−8223.

(14) Michalak, A. M.; Anderson, E. J.; Beletsky, D.; Boland, S.; Bosch, N. S.; Bridgeman, T. B.; Chaffin, J. D.; Cho, K.; Confesor, R.; Daloğlu, I.; DePinto, J. V.; Evans, M. A.; Fahnenstiel, G. L.; He, L.; Ho, J. C.; Jenkins, L.; Johengen, T. H.; Kuo, K. C.; LaPorte, E.; Liu, X.; McWilliams, M. R.; Moore, M. R.; Posselt, D. J.; Richards, R. P.; Scavia, D.; Steiner, A. L.; Verhamme, E.; Wright, D. M.; Zagorski, M. A. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (16), 6448−6452.
(15) Rabalais, N. N.; Turner, R. E. Gulf of Mexico hypoxia: past, present, and future. *Limnol. Oceanogr. Bull.* **2019**, *28* (4), 117−124.
(16) Jones, C. S.; Nielsen, J. K.; Schilling, K. E.; Weber, L. J. Iowa stream nitrate and the Gulf of Mexico. *PLoS One* **2018**, *13* (4), No. e0195930.
(17) Petrolia, D. R.; Gowda, P. H. Missing the Boat: Midwest Farm Drainage and Gulf of Mexico Hypoxia. *Appl. Econ. Perspect. Pol.* **2006**, *28* (2), 240−253.
(18) Turner, R. E.; Rabalais, N. N. Changes in Mississippi River water quality this century. *BioScience* **1991**, *41*, 140−147.
(19) Crawford, J. T.; Stets, E. G.; Sprague, L. A. Network controls on mean and variance of nitrate loads from the Mississippi River to the Gulf of Mexico. *J. Environ. Manage.* **2019**, *48*, 1789−1799.
(20) 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* (10), 1739−1745.
(21) 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* (5), No. e1373.
(22) Schindler, D. W.; Hecky, R. E.; Findlay, D. L.; Stainton, M. P.; Parker, B. R.; Paterson, M. J.; Beaty, K. G.; Lyng, M.; Kasian, S. E. M. Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (32), 11254−11258.
(23) National Science and Technology Council, Subcommittee on Ocean Science and Technology. *Harmful Algal Blooms and Hypoxia Comprehensive Research Plan and Action Strategy: An Interagency Report*, 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/NSTC/final_habs_hypoxia_research_plan_and_action.pdf.
(24) Paerl, H. W. Mitigating harmful cyanobacterial blooms in a human- and climatically-impacted world. *Life* **2014**, *4*, 988−1012.
(25) Liu, W.; Li, S.; Bu, H.; Zhang, Q.; Liu, G. Eutrophication in the Yunnan Plateau lakes: the influence of lake morphology, watershed land use, and socioeconomic factors. *Environ. Sci. Pollut. Res.* **2012**, *19*, 858−870.
(26) Reichwaldt, E. S.; Ghadouani, A. Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: Between simplistic scenarios and complex dynamics. *Water Res.* **2012**, *46* (5), 1372−1393.
(27) Ranjbar, M. H.; Hamilton, D. P.; Etemad-Shahidi, A.; Helfer, F. Impacts of atmospheric stilling and climate warming on cyanobacterial blooms: An individual-based modelling approach. *Water Res.* **2022**, *221*, 118814.
(28) Chaffin, J. D.; Davis, T. W.; Smith, D. J.; Baer, M. M.; Dick, G. J. Interactions between nitrogen form, loading rate, and light intensity on *Microcystis* and *Planktothrix* growth and microcystin production. *Harmful Algae* **2018**, *73*, 84−97.
(29) Cao, C.; Zheng, B.; Chen, Z.; Huang, M.; Zhang, J. Eutrophication and algal blooms in channel type reservoirs: A novel enclosure experiment by changing light intensity. *J. Environ. Sci.* **2011**, *23*, 1660−1670.
(30) Huang, J.; Xu, Q.; Wang, X.; Ji, H.; Quigley, E. J.; Sharbatmaleki, M.; Li, S.; Xi, B.; Sun, B.; Li, C. Effects of hydrological and climatic variables on cyanobacterial blooms in four large shallow lakes fed by the Yangtze River. *Environ. Sci. Ecotechnology* **2021**, *5*, 100069.
(31) Huo, S.; He, Z.; Ma, C.; Zhang, H.; Xi, B.; Xia, X.; Xu, Y.; Wu, F. Stricter nutrient criteria are required to mitigate the impact of climate change on harmful cyanobacterial blooms. *J. Hydrol.* **2019**, *569*, 698−704.
(32) Chong, S.; Lee, H.; An, K. G. Predicting taste and odor compounds in a shallow reservoir using a three-dimensional hydrodynamic ecological model. *Water* **2018**, *10* (10), 1396.
(33) Lee, S.; Lee, D. Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. *Int. J. Environ. Res. Public Health* **2018**, *15* (7), 1322.
(34) Wang, H.; Zhu, R.; Zhang, J.; Ni, L.; Shen, H.; Xie, P. A novel and convenient method for early warning of algal cell density by chlorophyll fluorescence parameters and its application in a highland lake. *Front. Plant Sci.* **2018**, *9*, 869.
(35) Wilkinson, G. M.; Carpenter, S. R.; Cole, J. J.; Pace, M. L.; Batt, R. D.; Buelo, C. D.; Kurtzweil, J. T. Early warning signals precede cyanobacterial blooms in multiple whole-lake experiments. *Ecol. Monogr.* **2018**, *88* (2), 188−203.
(36) Kim, T. H.; Shin, J.; Lee, D. Y.; Kim, Y. W.; Na, E.; Park, J. H.; Lim, C.; Cha, Y. K. Simultaneous feature engineering and interpretation: Forecasting harmful algal blooms using a deep learning approach. *Water Res.* **2022**, *215*, 118289.
(37) Ahn, J. M.; Kim, J.; Park, L. J.; Jeon, J.; Jong, J.; Min, J. H.; Kang, T. Predicting cyanobacterial harmful algal blooms (Cyano-HABs) in a regulated river using a revised EFDC model. *Water* **2021**, *13* (4), 439.
(38) Li, H.; Qin, C.; He, W.; Sun, F.; Du, P. Improved predictive performance of cyanobacterial blooms using a hybrid statistical and deep-learning method. *Environ. Res. Lett.* **2021**, *16*, 124045.
(39) Myer, M. H.; Urquhart, E.; Schaeffer, B. A.; Johnston, J. M. Spatio-Temporal Modeling for Forecasting High-Risk Freshwater Cyanobacterial Harmful Algal Blooms in Florida. *Front. Environ. Sci.* **2020**, *8*, 581091.
(40) Pyo, J. C.; Park, L. J.; Pachepsky, Y.; Baek, S. S.; Kim, K.; Cho, K. H. Using convolutional neural network for predicting cyanobacteria concentrations in river water. *Water Res.* **2020**, *186*, 116349.
(41) Thomas, M. K.; Fontana, S.; Reyes, M.; Kehoe, M.; Pomati, F. The predictability of a lake phytoplankton community, over time-scales of hours to years. *Ecol. Lett.* **2018**, *21* (5), 619−628.
(42) Chapra, S. C.; Boehlert, B.; Fant, C.; Bierman, V. J., Jr.; Henderson, J.; Mills, D.; Mas, D. M. L.; Rennels, L.; Jantarasami, L.; Martinich, J.; Strzepek, K. M.; Paerl, H. W. Climate change impacts on harmful algal blooms in U.S. freshwaters: A screening-level assessment. *Environ. Sci. Technol.* **2017**, *51* (16), 8933−8943.
(43) Kerimoglu, O.; Jacquet, S.; Vinçon-Leite, B.; Lemaire, B. J.; Rimet, F.; Soulignac, F.; Trévisan, D.; Anneville, O. Modelling the plankton groups of the deep, peri-alpine Lake Bourget. *Ecol. Model.* **2017**, *359*, 415−433.
(44) Recknagel, F.; Orr, P. T.; Bartkow, M.; Swanepoel, A.; Cao, H. Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling. *Harmful Algae* **2017**, *69*, 18−27.
(45) de J Magalhães, A. A.; da Luz, L. D.; de Aguiar Junior, T. R. Environmental factors driving the dominance of the harmful bloom-forming cyanobacteria *Microcystis* and *Aphanocapsa* in a tropical water supply reservoir. *Water Environ. Res.* **2019**, *91* (11), 1466−1478.
(46) Bui, M.-H.; Pham, T.-L.; Dao, T.-S. Prediction of cyanobacterial blooms in the Dau Tieng Reservoir using an artificial neural network. *Mar. Freshwater Res.* **2017**, *68* (11), 2070−2080.
(47) Tyler, A. N.; Hunter, P. D.; Carvalho, L.; Codd, G. A.; Elliott, J. A.; Ferguson, C. A.; Hanley, N. D.; Hopkins, D. W.; Maberly, S. C.; Mearns, K. J.; Scott, E. M. Strategies for monitoring and managing mass populations of toxic cyanobacteria in recreational waters: a multi-interdisciplinary approach. *Environ. Health* **2009**, *8*, S11.
(48) Recknagel, F.; Branco, C. W. C.; Cao, H.; Huszar, V. L. M.; Sousa-Filho, I. F. Modelling and forecasting the heterogeneous distribution of picocyanobacteria in the tropical Lajes Reservoir (Brazil) by evolutionary computation. *Hydrobiologia* **2015**, *749* (1), 53−67.


https://doi.org/10.1021/acs.est.3c07764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(49) Cao, H.; Recknagel, F.; Bartkow, M. Spatially-explicit forecasting of cyanobacteria assemblages in freshwater lakes by multi-objective hybrid evolutionary algorithms. *Ecol. Model.* **2016**, *342*, 97−112.
(50) Recknagel, F.; Orr, P. T.; Cao, H. Inductive reasoning and forecasting of population dynamics of *Cylindrospermopsis raciborskii* in three sub-tropical reservoirs by evolutionary computation. *Harmful Algae* **2014**, *31*, 26−34.
(51) Mchau, G. J.; Makule, E.; Machunda, R.; Gong, Y. Y.; Kimanya, M. Phycocyanin as a proxy for algal blooms in surface waters: case study of Ukerewe Island, Tanzania. *Water Pract. Technol.* **2019**, *14* (1), 229−239.
(52) Li, W.; Qin, B. Dynamics of spatiotemporal heterogeneity of cyanobacterial blooms in large eutrophic Lake Taihu, China. *Hydrobiologia* **2019**, *833* (1), 81−93.
(53) Kasinak, J. M. E.; Holt, B. M.; Chislock, M. F.; Wilson, A. E. Benchtop fluorometry of phycocyanin as a rapid approach for estimating cyanobacterial biovolume. *J. Plankton Res.* **2015**, *37* (1), 248−257.
(54) *The Iowa Department of Natural Resources Water Quality Monitoring and Assessment (AQuIA)*, 2023. https://programs.iowadnr.gov/aquia/Programs/Beaches (accessed Sept, 2023).
(55) Lee, J.; Choi, J.; Fatka, M.; Swanner, E.; Ikuma, K.; Liang, X.; Leung, T.; Howe, A. Improved detection of *mcy*A genes and their phylogenetic origins in harmful algal blooms. *Water Res.* **2020**, *176*, 115730.
(56) Muyzer, G.; de Waal, E. C.; Uitterlinden, A. G. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* **1993**, *59* (3), 695−700.
(57) Ritalahti, K. M.; Amos, B. K.; Sung, Y.; Wu, Q.; Koenigsberg, S. S.; Löffler, F. E. Quantitative PCR targeting 16S rRNA and reductive dehalogenase genes simultaneously monitors multiple *Dehalococcoides* strains. *Appl. Environ. Microbiol.* **2006**, *72* (4), 2765−2774.
(58) U.S. Geological Survey. *National Land Cover Database (NLCD) 2019 Products (Ver. 2.0, June 2021)*, 2021. (accessed March, 2023).
(59) *Raster: Geographic Analysis and Modeling*. http://CRAN.R-project.org/package=raster (accessed March, 2023).
(60) Pebesma, E. Simple features for R: Standardized support for spatial vector data. *R J.* **2018**, *10* (1), 439−446.
(61) *Exactextractr: Fast Extraction from Raster Datasets Using Polygons*. https://CRAN.R-project.org/package=exactextractr (accessed March, 2023).
(62) *Weather Underground*, 2023. https://www.wunderground.com/ (accessed March, 2023).
(63) Lee, J.; Im, J.; Kim, U.; Löffler, F. E. A data mining approach to predict in situ detoxification potential of chlorinated ethenes. *Environ. Sci. Technol.* **2016**, *50* (10), 5181−5188.
(64) *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. https://www.tidymodels.org (accessed March, 2023).
(65) Friedman, J. H.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* **2010**, *33* (1), 1−22.
(66) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016; pp 785−794.
(67) Kuhn, M.; Falbel, D. *Brulee: High-Level Modeling Functions with "Torch"*. https://github.com/tidymodels/brulee (accessed March, 2023).
(68) Ranganathan, P.; Aggarwal, R. Understanding the properties of diagnostic tests − Part 2: Likelihood ratios. *Perspect. Clin. Res.* **2018**, *9* (2), 99−102.
(69) Tillett, D.; Dittmann, E.; Erhard, M.; von Döhren, H.; Börner, T.; Neilan, B. A. Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: an integrated peptide−polyketide synthetase system. *Chem. Biol.* **2000**, *7* (10), 753−764.
(70) Hu, L.; Shan, K.; Huang, L.; Li, Y.; Zhao, L.; Zhou, Q.; Song, L. Environmental factors associated with cyanobacterial assemblages in a mesotrophic subtropical plateau lake: A focus on bloom toxicity. *Sci. Total Environ.* **2021**, *777*, 146052.
(71) Miao, K.; Li, X.; Guo, L.; Gao, M.; Zhao, Y.; Jin, C.; Ji, J.; She, Z. Cultivation of *Chlorella pyrenoidosa* with different phosphorus forms under photoautotrophic and mixotrophic modes: Biochemical component synthesis and phosphorus bioavailability appraisement. *J. Clean. Prod.* **2022**, *359*, 132058.
(72) Horecka, H. M. Environmental factors linked to harmful algal bloom induced shellfish toxicity in Cobscook Bay, Maine. Honors Thesis, Honors College, 2012. https://digitalcommons.library.umaine.edu/honors/56.
(73) Leigh, C.; Burford, M. A.; Roberts, D. T.; Udy, J. W. Predicting the vulnerability of reservoirs to poor water quality and cyanobacterial blooms. *Water Res.* **2010**, *44* (15), 4487−4496.
(74) Dueben, P. D.; Bauer, P. Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.* **2018**, *11*, 3999−4009.
(75) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.; Zhang, H. Machine learning: new ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741−12754.
(76) Moosavi, V.; Gheisoori Fard, Z.; Vafakhah, M. Which one is more important in daily runoff forecasting using data driven models: Input data, model type, preprocessing or data length? *J. Hydrol.* **2022**, *606*, 127429.
(77) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N.; Baker, N. How much chemistry does a deep neural network need to know to make accurate predictions?. *2018 IEEE Winter Conference on Applications of Computer Vision*, 2018; pp 1340−1349.


https://doi.org/10.1021/acs.est.3c07764
*Environ. Sci. Technol.* XXXX, XXX, XXX−XXX