JOURNAL OF
EMERGING INVESTIGATORS

Article

# Using two-step machine learning to predict harmful algal bloom risk

**Aryaman Shukla[1], Ravi Shukla[1]**

[1] Simon G. Atkins Academic & Technology High School, Winston Salem, North Carolina

## SUMMARY

**Water shortages are a global issue now impacting North America. Inland water quality is significantly affected by the seasonal occurrence of harmful algal blooms (HABs). Annually, the economic impact of HAB is over $2.3 billion due to the cleanup, drinking water restrictions, tourism, and closure of fisheries. Existing machine learning (ML) models use binary classifications, such as the presence or absence of HABs, to predict cyanobacteria proliferation, which can leave a gap in assessing the likelihood of a potential outbreak. In this study, we explored the application of ML regression algorithms to predict HAB risk on a continuum. Using primary data from water samples collected in Maryland, North Carolina, and Virginia, we hypothesized there would be a positive correlation between algal weight (an indicator of HAB risk) and nitrates, phosphates, and temperature. To test this hypothesis, we trained artificial intelligence (AI) models using primary data collected from 30 inland aquatic systems. Using the results, we then built Monte Carlo simulations generating over 100,000 scenarios to perform sensitivity analysis on the variables to predict the HAB risk. Through our experiments and selecting ML regression models with high validation accuracies, we achieved a 77% test accuracy in predicting HAB risk levels. We checked the results of our HAB forecasts with observations from United States Geological Survey (USGS) and National Aeronautics and Space Administration (NASA) for specific locations and dates, further validating the model's accuracy.**

## INTRODUCTION

Water contamination is a global issue that has also begun affecting North America (1). Over the past decade, New York has seen an almost twenty-fold increase in harmful algal blooms (HABs) reported by the New York State Department of Environmental Conversation (2). However, this is not limited to New York alone. These seasonal outbreaks of HABs during the summer months, exacerbated by climate change, have also affected the availability of clean drinking water in California (3). Furthermore, nutrient loading has increased the incidents of HABs in regions previously perceived as less affected (1). Frequent, intense blooms with longer seasons and expanded geographic distribution driven by climate change, nutrient pollution, and evolutionary pressures are further aggravated by environmental and ecological pressures that select more toxic strains of HABs, chiefly cyanobacteria (4). As temperatures rise, slow-moving water experiences thermal stratification, creating a warm layer conducive to algal blooms atop cooler water. This enhances cyanobacteria's buoyancy and downstream dispersal, concentrating toxin-laden mats along benthic shorelines and increasing public health risks to humans and animals (5, 6). The spread of microcystin-producing cyanobacteria HABs (cyanoHABs) from inland to marine waters underscores their impact on marine life and potential human exposure via seafood consumption (7). Cyanobacteria remain viable and capable of maintaining their toxicity after being released from freshwater systems downstream (8). From eutrophication alone, the annual economic impact of freshwater HABs is estimated to be $2.2 billion (4).

The increasing recognition of HABs in both lentic (standing) and lotic (flowing) inland waters as natural phenomena has spurred efforts for early detection and monitoring (9). For inland water management, authorities have employed tailored local strategies to monitor water quality. Early-warning systems (EWSs) are based on fixed-location sample collection and analysis. However, financial and staffing limitations results in reduced geographic coverage and irregular sampling (10). Furthermore, the time lag in acquiring analytical results and the limited interpretative abilities diminish the response effectiveness in prediction, prevention, and mitigation for authorities overseeing drinking water distribution and recreational water use (10).

Another approach to monitoring the cyanoHABs is the use of remote sensing, which utilizes the electromagnetic spectrum via satellite-derived values. These efforts involve the use of satellite images and data collected based on specific wavelengths for indicator pigments like chlorophyll-a and phycocyanin (9, 11). While satellite imaging has proven effective for mapping large bodies of water like oceans, coasts, and major lakes, inland waters require higher-resolution spatial data, which is scarce. Additionally, HAB detection in turbid waters poses challenges as turbidity interferes with the detection of chlorophyll-a and phycocyanin (9).

For inland waters, agencies mitigate risks through an integrated machine learning (ML) model, enabling effective contingency plans. However, the existing ML models either rely on large, labeled datasets that are localized, or are limited to a single body of water. As they are classification ML models, they typically employ a binary approach to predict the presence or absence of HABs (12, 13). Due to the localized nature of these models and the lack of standardized and coordinated efforts across the US, it is challenging to calibrate routine water quality measuring projects with remotely sensed Earth observation data (9,11). Combining resources to address HAB outbreaks is also difficult logistically. Therefore, there is a noticeable need for a standardized reporting method (9). Multi-factor data driven models have been used in recent

https://doi.org/10.59720/24-196

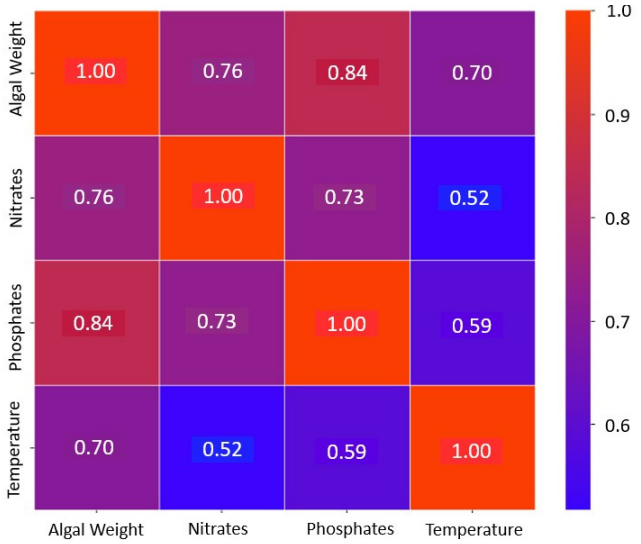


**Figure 1: Correlation matrix of the algal weight with temperature, nitrates, and phosphates.** The color of the box reflects the strength of the correlation with blue reflecting a weak correlation and red a perfect correlation. The correlation was run on all 240 samples that were collected in person from bodies of water in NC, MD, and VA. A strong correlation was observed between the dependent variable (algal weight), and the independent variables (nitrates, phosphates, and temperature).

years to predict cyanobacteria blooms (14). This creates an opportunity for a multi-state data driven model that can establish parameters to gauge the risk of HAB occurrence in a given water body.

To predict HAB threat levels, we employed classification and regression ML models based on primary data to identify the variables correlated with HAB presence and the associated coefficients. These coefficients were then applied to Monte Carlo models to simulate HAB threat levels across the mid-Atlantic US. We hypothesized that there would be a positive correlation between algal weight (an indicator of HAB risk), nitrate and phosphate levels in the water, and the external temperature. We also hypothesized that these independent variables, nitrate, phosphate, and temperature, would be significant predictors of algal weight and that the ML model's coefficients could quantify the impact that a percentage change in the independent variables would have on algal weight. With this approach, our model would provide the risk of HAB on a continuum rather than as a binary classification of presence or absence of HAB. Using primary data collected from 30 sites across three mid-Atlantic states, we addressed the complexity of simulating HAB growth dynamics by integrating ML regression models into Monte Carlo unweighted stochastic sampling simulations, which provided sensitivity analysis highlighting the impact of an individual variable on algal growth. We further corroborated the predicted results of the ML models by referencing the USGS and NASA observations for specific sites for the same time periods as the primary data. Our results showed a strong relationship between the algal weight and the levels of nitrates, phosphates, and temperature. The ML model could explain over 77% of the variation in the algal weight and from it the likelihood of HAB presence based on the available readings of nitrates, phosphates, and temperature. By extrapolating these readings, the model could also predict when an HAB occurrence could emerge.

## RESULTS

We first obtained primary data that provided readings on the algal weight as well as the levels of nitrate and phosphate in the water and the water temperature to model the impact of these factors on HAB risk. We collected water samples from three states over a five-month period to ensure that the data captured the variables and algal weight over time and represented a large area. The 240 water samples collected were from 30 inland waterbodies across North Carolina, Virginia, and Maryland. Additionally, we used rainwater samples collected from five locations in proximity to the test sites as a control to check for levels of nitrates and phosphates in natural, uncontaminated water. We observed a correlation between the nitrate, phosphate, and temperature levels and the algal weight by generating a heatmap of the correlation matrix, which showed a positive correlation between the algal weight and the nitrates (0.76), phosphates (0.84) and temperature (0.70) in the water, confirming their relevance to the model **(Figure 1)**. Using the correlation matrix we also examined nitrate, phosphate, and temperature for the presence of high correlation that could indicate a risk of multicollinearity. Typically, correlation values above 0.80 between pairs of independent variables suggest that potential multicollinearity may be present. This was not observed in the correlation matrix. Our water samples indicated that algal presence intensified at temperatures between 70 and 90°F. **(Figure 2)**.

The relationship between algal weight and HAB risk was

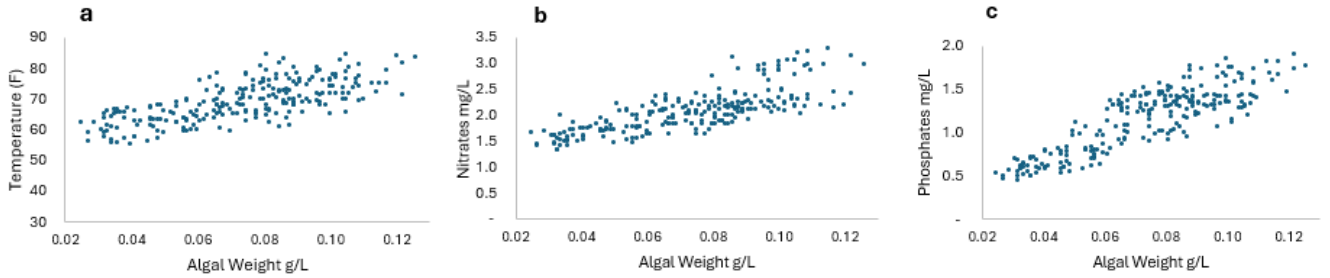


**Figure 2: Scatter plots of the interaction between the dependent (algal weight) and independent (temperature, nitrates, phosphates) variables.** The visual shows that the algal weight trends upwards and has a positive correlation with the temperature (a), nitrate (b), and phosphate (c) levels. Each dot represents one of 240 water samples collected and the algal weight, nitrate, phosphate, and temperature values associated with it.

https://doi.org/10.59720/24-196

established by performing an ELISA test on the samples collected from each site. To quantify the risk, we identified bands of algal weight and the HAB presence (risks) associated with them using the collected water samples. Based on the microcystins levels observed in the collected samples, the low eutrophication band had algal weight < 0.06gm/L, moderate eutrophication band had algal weight > 0.06 and < 0.10 gm/L, and high eutrophication band had algal weight > 0.10gm/L. The low eutrophication bands had low indication of HAB presence while the high eutrophication bands usually had a confirmed HAB presence. Thus, by forecasting the algal weight, our model would be able to predict the HAB risk associated with that level of algal weight.

We conducted a regression analysis individually for each state, using the state-specific data, to see whether the relationship between nitrate and phosphate levels and algal weight was consistent across all three states. The regression coefficients were similar across states, which allowed us to aggregate all data into a single database. This facilitated the development of a comprehensive ML model for the Mid-Atlantic region. To ensure there was a diversity of techniques and a robust comparison, a mix of linear, nonlinear, ensemble and decision tree models were selected. We tested multiple ML regression models that are widely used and recognized for their application in regression problems to confirm the relationship between the algal weight and the levels of nitrates, phosphates, and temperature. The models were trained using 70% of the data and their performance was tested on the remaining 30%.

Results from the ML regression models showed the polynomial regression and linear regression models performed the best, with both having an $R^2$ of 0.77. This indicates that by using the three independent variables—nitrates, phosphates, and temperature—the polynomial and linear regression models explained 77% of the change in algal weight. Further, the relationship between the algal weight, and the nitrates, phosphates, and temperature readings was found to be significant ($p < 0.05$). The $R^2$ values for the other models were between 0.65 and 0.74 **(Figure 3)**.

Once the relationships between the algal weight, nitrates, phosphates, and temperature had been quantified, we used the coefficients associated with these variables from the linear regression model to understand the contribution of each variable in driving the change in algal weight. To estimate this, we built a Monte Carlo simulation model that enabled us to calculate the impact of a percentage change in each independent variable on the algal weight and visualized the results in a histogram. The addition of the Monte Carlo simulations served two purposes. Given that the original data was limited to 240 observations, by taking the mean and standard deviation of the variables and applying the model coefficients to run stochastic sampling simulations, we were able to use the probability distributions to generate an accurate prediction of the algal weight and HAB risk for any set of data. We also used the Monte Carlo simulations to run a sensitivity analysis to accurately predict the impact of the change in the independent variables on the HAB risk by creating a representative sample of outcomes using thousands of possible combinations.

Our simulations indicated that a concurrent 1% rise in nitrates, phosphates, and temperature could increase HAB presence by 33% from the current levels. A 5% increase in these variables revealed that the presence of HAB would surge from 4.9%, the HAB level seen in our sampled water bodies at the time of sampling, to 13.0%, an increase of almost 160% **(Figure 4)**. On the other hand, we found that a 5% decrease in the nitrate, phosphate, and temperature levels would increase the number of low HAB-risk water bodies at the time of sampling from 20% to 36.9%, an increase of 86%, while simultaneously the high HAB risk waterbodies would reduce by 76% from 5% to 1.2% **(Figure 5)**. These results indicate that controlling the levels of nitrates and phosphates might be a tangible solution to reduce HAB risk.

Next, we wanted to compare our model's predictions to real

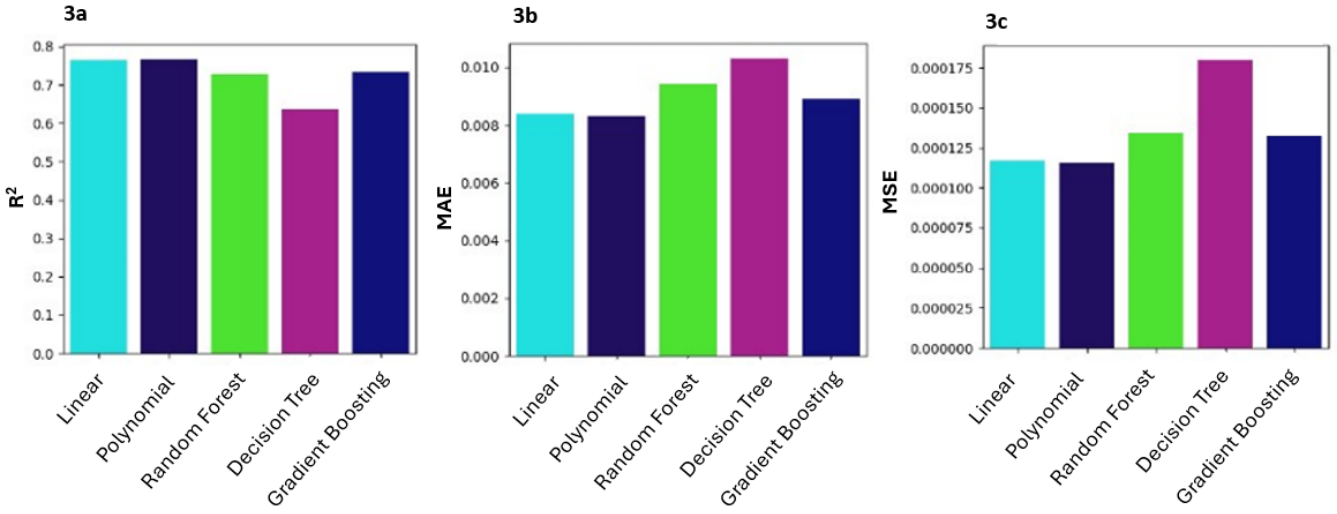


**Figure 3: ML model performance predicting algal weight from nitrate, phosphate, and temperature levels.** Results comparing the Coefficent of Determination ($R^2$), the Mean Absolute Error (MAE), and the Mean Squared Error (MSE) of five ML models using a 70:30 train:test split of 240 water samples. The polynomial and linear regression models had the highest $R^2$ (a) and the lowest MAE (b) and MSE (c) and performed the best, followed by Gradient Boosting model. The $R^2$ reflects the proportion of the variation in the algal weight that can be predicted from the nitrate, phosphate, and temperature levels.

https://doi.org/10.59720/24-196

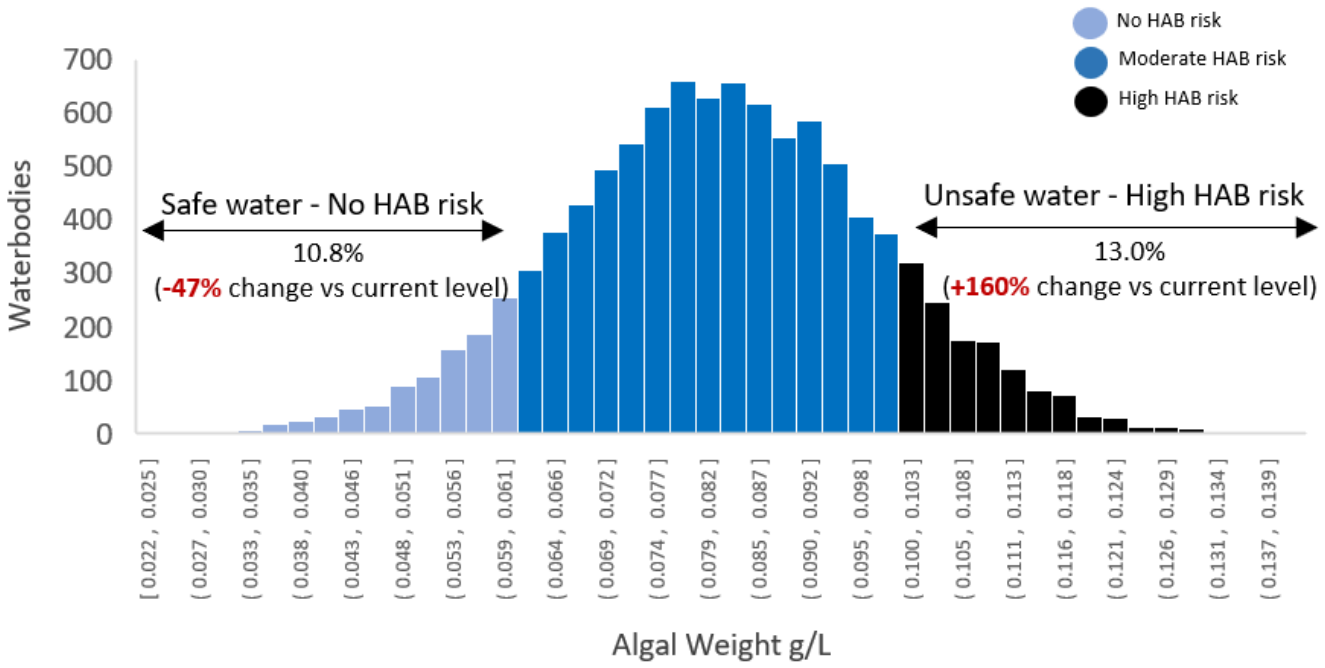


**Figure 4: Monte Carlo model predictions from 10,000 simulations showing the impact on HAB risk from a 5% increase in nitrate, phosphate, and temperature levels.** Results show the impact on water quality and HAB risk due to a 5% increase in the nitrates, phosphates, and temperature levels. The percentage of low HAB-risk water bodies halved from ~20% to 10.8% while the percentage of high HAB-risk water bodies increased from ~5% to 13.0%.

world data from other bodies of water to confirm its accuracy. From the USGS water quality portal, we downloaded the data for all sites and selected those in NC, MD, and VA where longitudinal data were available on the nitrate, phosphate, and temperature readings. The data readings on the predictors—nitrate, phosphate, and temperature levels—are frequently collected by USGS for numerous locations across the country on a weekly or monthly frequency (15). Using the coefficients, from the linear regression model we ran our calculations to estimate the algal weight in these locations and used it to predict the risk of a HAB presence. Based on the presence of HAB, we could also identify the HAB risk for each body of water over time **(Figure 7)**. We validated our model results on the risk of HAB for sampled sites with data available from the Department of Environment websites of each state, as well as the National Oceanic and Atmospheric Administration (NOAA) (22-24). The results from our two-step model aligned with the reports published by the states and NOAA for these locations and periods. For example, using the USGS data, in June 2023, a high HAB presence was predicted by our model and confirmed by the Maryland state department website, *eyesonthebay*, at Choptank River in MD (22). During the same period, the USGS data indicated that Plumtree Run in MD had algal presence in the low-risk band while for Contentnea Creek and Neuse River in NC, the data showed a slowly increasing risk in the moderate zone These predictions could be validated from the Maryland and North Carolina state websites. Using the widely available USGS data in the model enables predicting the HAB risk on a continuum compared to the existing binary models that only identify the presence or absence of HAB, which facilitates prioritizing resources to areas where the risk is the highest.

## DISCUSSION

Constructing a regression model that can predict algal weight, a leading indicator of HAB risk, is significant not only because the model identifies the variables that drive HAB risk but also because it allows the risk to be measured on a continuum, which provides valuable insight for mitigation efforts as the risk approaches a high level. Our findings of a strong correlation between the algal weight and nitrate, phosphate, and temperature levels help to deepen the understanding of the impact these variables have on algal weight, a predictor of HAB risk. This understanding makes it easier to quantify the risk that increasing levels of nitrates, phosphates, and temperature pose in promoting HAB presence and facilitates developing targeted solutions to mitigate this risk.

The $R^2$ or the coefficient of determination represents the change in the dependent variable that is explained by the independent variables. Since we set the $R^2 > 0.75$ as the benchmark for our model, taking into consideration the natural variability of the measured variables, achieving an $R^2$ value of 0.77 may seem modest; however, it reflects a realistic assessment of model performance given the complexity of predicting algal weight. Several factors could contribute to why the $R^2$ value was not higher. First, the inherent variability in biological systems like algal growth can be influenced by numerous, sometimes unpredictable environmental factors, which are not always fully captured by the model's feature variables. The unexplained loss of 0.23 in the $R^2$ value can also indicate geographical variations across the different states, recognizing that algal growth is subject to a wide range of region-specific environmental influences that may not be uniformly represented in the model. Additionally,

https://doi.org/10.59720/24-196

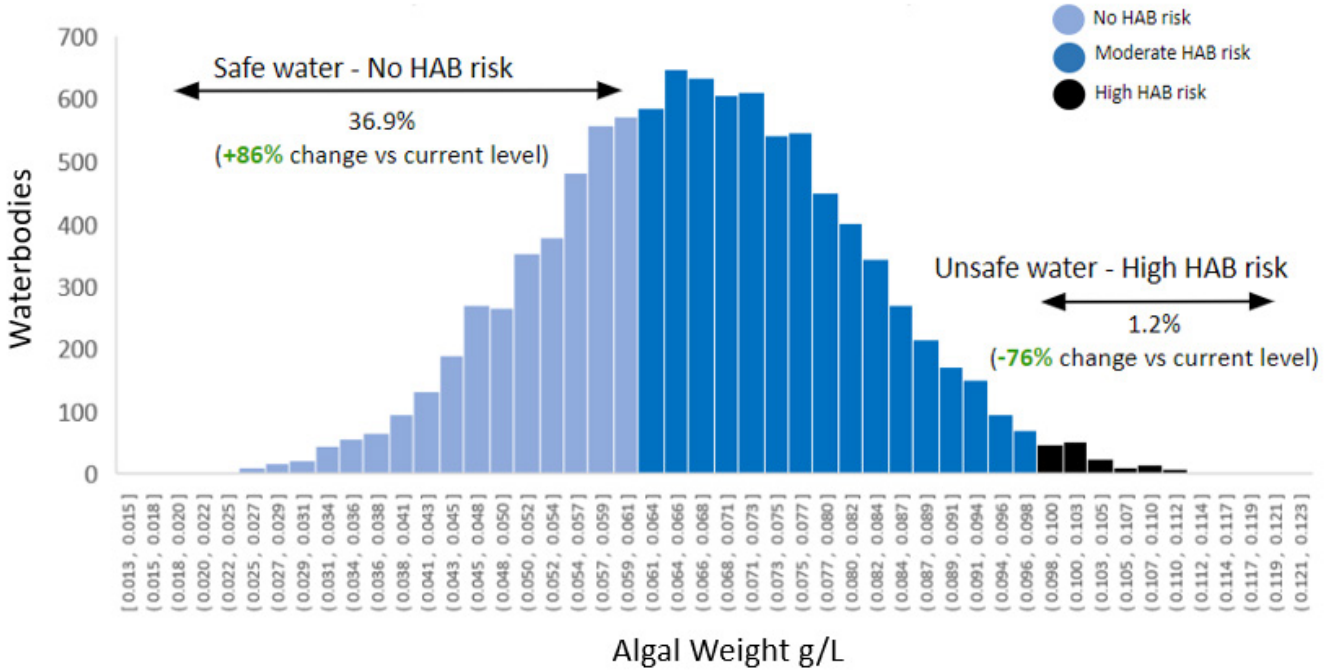


**Figure 5: Monte Carlo model predictions from 10,000 simulations showing the impact on HAB risk from a 5% decrease in nitrate, phosphate, and temperature levels.** Results show the impact on water quality and HAB risk due to a 5% decrease in nitrates, phosphates, and temperature levels. The percentage of low HAB-risk water bodies jumps from ~20% to 36.9%. In comparison, the percentage of high HAB-risk water bodies significantly drops from ~5% to 1.2%, highlighting the positive impact we could have by reducing the eutrophication of our waterbodies.

the relationship between these variables and algal weight might not be linear or polynomial, leading to some degree of unexplained variance. Finally, measurement errors or limitations in the data quality and quantity can also constrain the $R^2$ value. These considerations highlight the challenges in modeling biological phenomena and the importance of setting realistic expectations for model accuracy.

The most significant challenge we faced was the limited availability of data from satellite images to corroborate our two-step model. The data predominantly focused on coastal and large open waters, which substantially restricted our ability to validate the model for inland regions using spectral images. However, as technology progresses, there is potential for more integrated data collection and model application across the U.S. This advancement is particularly crucial for diverse inland water systems, where environmental conditions can differ markedly from those in coastal areas. In such contexts, ML models become increasingly valuable predictive tools, enabling preemptive actions to safeguard water quality.

Given the constraints in resources and time, we developed the models using a dataset of 240 primary samples collected through direct fieldwork. Despite the limited dataset, the constructed ML models not only fulfilled but surpassed our predefined benchmarks. However, we can elevate the predictive efficacy of these models by incorporating an expanded dataset. Initially, we employed an 80-20 data split for training and testing purposes, to maximize use of the training set. We subsequently revised the approach to a 70-30 split, motivated by the need to enhance the robustness of the testing process. With this recalibration, we aimed to achieve an optimal balance to mitigate the risk of overfitting while ensuring adequate model validation. Looking forward, the iterative addition of new data points and periodic updating of the model are expected to yield continuous improvements. Specifically, we expect that such updates will lead to a refinement of the model and further contribute to a greater level of accuracy in predictions. This evolving process underscores the dynamic nature of ML models, where ongoing data integration plays a pivotal role in enhancing model performance and reliability.

Utilizing the two-step model developed in this study, we may be better able to predict the severity of HAB outbreaks based on various factors sourced from the USGS of the National Water Information System (NWIS), thereby

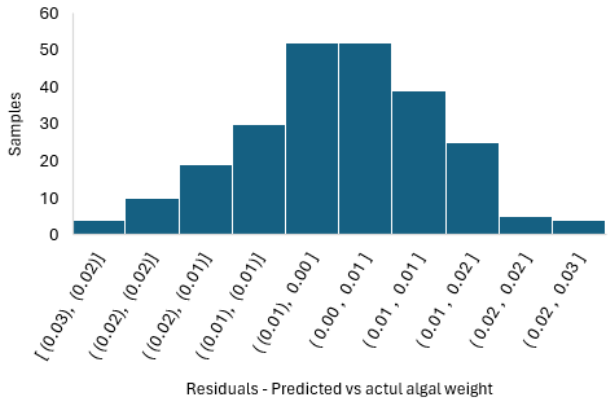


**Figure 6: Histogram of residuals from the linear regression model.** A visual presentation of the residuals of the actual algal weight from the samples and predicted algal weight generated by the linear regression model shows a normal distribution.

https://doi.org/10.59720/24-196

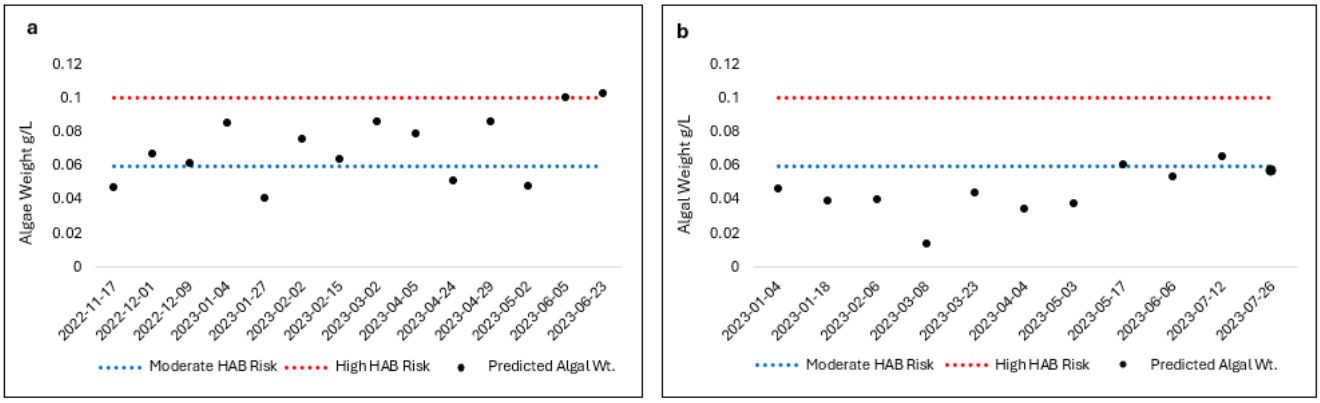


**Figure 7: Model predictions of HAB presence in two sample water bodies on a time and risk continuum using USGS data feed.** The nitrate, phosphate, and temperature levels from the USGS data feed for different locations were fed into the model. The blue line shows the algal weight where the risk of HAB presence changes from low to moderate, the red line indicates the level where the risk changes from moderate to high. The model indicates that over the time period, based on the levels of the independent variables and the predicted algal weight, the HAB risk at Choptank River has been between moderate to high (a). By contrast the HAB risk in the Neuse River was low to moderate (b). The data range is not continuous.

adding incremental value to an existing resource that is left underutilized (15). This unified approach contrasts with the current reliance on varying state-level definitions and thresholds for HAB risk, making it challenging to implement a cohesive strategy. For instance, North Carolina defines HAB risk based on visible discoloration or surface scum, while New York sets Microcystin-LR thresholds at > 4 µg/L, and Virginia defines Microcystin risk at > 6 µg/L (16). Additionally, achieving effective mitigation requires balancing strategies that are not just focused on minimizing environmental impacts but also on mitigating costs. As these strategies evolve to meet stakeholder requirements, an understanding of the impact each variable has on HAB risk becomes even more relevant (17).

Although the model serves as an important tool, it is crucial to continue studying HAB patterns and changes in toxicity levels as environmental scenarios evolve with climate change (18). Therefore, repeated experimentation and similar research are necessary to develop more comprehensive results. This research can serve as a foundation for future studies to use computational methods in the prediction of algal weight as an indicator of HAB risks, fostering a more unified approach to this environmental challenge.

## MATERIALS AND METHODS

### Sample Collection and Analysis

Over 5 months, we collected 4 sets of 500 mL grab samples from 30 inland lentic and lotic aquatic systems across NC, MD, and VA, with required safety protocols in place. During sample collection, we used Vernier probes to record specific parameters associated with HAB presence, including temperature, pH, and turbidity. The collected water samples were filtered, and the weight of the algal biomass was calculated after a 48-hour period. Additionally, we analyzed the grab samples for nitrate and phosphate content using Hach reagents, Nitrate TNT 835 and Phosphate TNT 843, and a spectrophotometer. For the control, we used rainwater samples collected from five sites, which determined the difference in the nitrate and phosphate content for analyses. Furthermore, we performed an ELISA test using the ABRAXIS Microcystins-ADDA for one sample per site to assess the presence and concentration of specific algae-related toxins, specifically microcystins and nodularins. This provided crucial information on the toxicity level of each HAB event and helped us identify the algal weight ranges associated with HAB risk. Based on the data comparing the algal weight with the microcystin readings, we created the following classification:

Algal weight < 0.06 gm/L = No risk (low eutrophication levels)
Algal weight > 0.06 and < 0.10 gm/L = Moderate risk (moderate eutrophication levels)
Algal weight > 0.10 gm/L = High risk (High eutrophication levels)

### Data Modeling

To conduct the correlation assessment, we referenced earlier studies that identified the key independent variables for HAB occurrence (17, 19, 20). Scatterplots were utilized to visualize the relationships between the dependent variable, algal weight, and the independent variables, including nitrates, phosphates, and temperature Pearson correlation coefficients were computed to quantify these associations. The sample data from each state was separately collated in Excel to run a regression analysis, and it was determined that all state data could be aggregated into a single database, which facilitated the development of a comprehensive ML model for the Mid-Atlantic United States.

The Pandas, NumPy, Seaborn, Matplotlib, and Scikit Learn libraries were used to analyze our dataset. In JupyterLab, we used the combined data to plot histograms. While the algal weight and temperature followed a normal distribution, the distribution for the nitrates and phosphates was a little skewed. This was noted and later checked against the distribution of the residuals to ensure the basic assumptions for running a linear regression were met **(Figure 6)**. Subsequently, we generated a correlation matrix to better understand the interaction between the algal weight, nitrates, phosphates, and temperature and confirm their relevance to the model. Additionally, we examined the presence of a high correlation among the feature variables, which could indicate a risk of

https://doi.org/10.59720/24-196

multicollinearity. We then constructed the models with Scikit Learn, which were trained using 70% of the data, and tested with the remaining 30%. To build the model, we assessed the performance of the five ML regression models—linear regression, polynomial regression, decision tree, random forest, and gradient boosting—using MSE, MAE, and R2 values. We established a minimum R2 value threshold of 0.75 for the models to be considered practical. The code of the models used in this research was programmed in Python 3.6 and is available on GitHub (21).

### Data Simulation

We next used the coefficients obtained for each variable from the ML regression algorithms in a Monte Carlo simulation. We organized the data in Excel by columns, arranging algal weight, nitrate, phosphate, and temperature was done to construct the Monte Carlo simulation. We then computed the mean and standard deviation for each variable from 240 data points. Subsequently, simulations were executed to generate values for each variable, utilizing the following formula: =NORM.INV(RAND(), MEAN, STDEV)

We multiplied the simulation results for the independent variables by the coefficients from the linear regression model to predict algal weight. These predictions were then displayed in a 10,000-row data table, generating 10,000 simulated values at a time. We conducted thousands of simulations, systematically adjusting predictors by varying percentages each time, and noted the change in the predicted values of the algal weight and their distribution in the low, medium, and high-risk classifications. Subsequently, we visualized the results in a histogram to categorize HAB presence into low, medium, and high levels. The output of 100,000 simulations was used to set a baseline.

To understand the impact of a percentage change in the independent variables, we increased the coefficients by 1%. The simulations were re-run to see the change in the predicted values of the algal weight and their distribution in the low, medium, and high classifications. We also ran similar simulations with 5% changes in the variables that reflected the impact of both negative and positive scenarios for the future. We took the results generated from 100,000 simulations to estimate the change in the high HAB prediction group. We also tested the robustness of the ML model's predictions in two ways. First, we input the USGS data from the National Water Quality portal that provides weekly or monthly data feed on water samples across different waterbodies in the US, including Maryland, North Carolina, and Virginia (15). This data included readings on nitrates, phosphates, and temperature, the variables used in our two-step model. We downloaded the data into an Excel file for all available sites in each state and selected those sites where longitudinal data was available for all three variables. We took data on nitrates, orthophosphates, and temperature readings from the weekly data feed. Using the coefficients, the model predicted the algal weight on a continuum, ranging from low HAB risk to high HAB risk.

Following this, we searched the USGS identification code to pinpoint the exact location for each reading. We then cross-checked them with the state environment monitoring and water management dashboards to confirm if HAB presence had been reported in these locations and days and compared the results with our model predictions. Secondly, we utilized the NOAA's National Centers for Coastal Ocean Science (NCCOS) Algal Bloom Beta Experimental Products website to access weekly cyanobacteria presence data recorded in the Chesapeake Bay area by the Sentinel-3A satellite mission (24). Information on cyanobacteria presence was employed to verify our modeled predictions for the weeks when in-person samples were collected in Maryland. Using the maps, we were able to gauge the level of cyanobacteria in a given area from the heat map index. Although we couldn't pinpoint the exact sampling location, the Sentinel-3A maps indicated the level of cyanobacteria presence within a one-mile area of the testing site on the day when the samples were collected.

## ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Kyana Young for her invaluable guidance, time, and expertise during the 2023 summer research assistant internship in her department. Dr. Young generously provided access to the Wake Forest University lab for testing. Her training in sample collection, observations, and testing laid the essential foundation for the successful execution of this project.

**Received:** August 3, 2024
**Accepted:** October 23, 2024
**Published:** JuLY 4, 2025

## REFERENCES

1. Smith, V.H. "Eutrophication." *Encyclopedia of Inland Waters*, 2009, pp. 61-73,
2. Wang, N., Mark, N., Launer, N., Hirtler, A., Weston, C., Cleckner, L., Faehndrich, C., LaGorga, L., Xia, L., Pyrek, D., Penningroth, S. M., & Richardson, R. E. (2024). Harmful algal blooms in Cayuga lake, NY: From microbiome analysis to eDNA monitoring. *Journal of Environmental Management*, *354*, 120128. https://doi.org/10.1016/j.jenvman.2024.120128
3. Solomon, Gina M., *et al.* "Notes From the Field: Harmful Algal Bloom Affecting Private Drinking Water Intakes — Clear Lake, California, June–November 2021." *MMWR Morbidity and Mortality Weekly Report*, vol. 71, no. 41, Oct. 2022, pp. 1306–07, https://doi.org/10.15585/mmwr.mm7141a3
4. Loftin, Keith A., *et al.* "Cyanotoxins in Inland Lakes of the United States: Occurrence and Potential Recreational Health Risks in the EPA National Lakes Assessment 2007." *Harmful Algae*, vol. 56, June 2016, pp. 77–90. https://doi.org/10.1016/j.hal.2016.04.001
5. Bouma-Gregson, Keith, *et al.* "Rise and Fall of Toxic Benthic Freshwater Cyanobacteria (Anabaena spp.) in the Eel River: Buoyancy and Dispersal." *Harmful Algae*, vol. 66, 2017, pp. 79-87. https://doi.org/10.1016/j.hal.2017.05.007
6. Gibble, C. M., and R. M. Kudela. "Detection of Persistent Microcystin Toxins at the Land–Sea Interface in Monterey Bay, California." *Harmful Algae*, vol. 39, 2014, pp. 146–153. https://doi.org/10.1016/j.hal.2014.07.004
7. Gibble, C. M., *et al.* "Evidence of Freshwater Algal Toxins in Marine Shellfish: Implications for Human and Aquatic Health." *Harmful Algae*, vol. 59, 2016, pp. 59–66. https://doi.org/10.1016/j.hal.2016.09.007
8. Howard, Meredith D. A., *et al.* "Integrative Monitoring

https://doi.org/10.59720/24-196

Strategy for Marine and Freshwater Harmful Algal Blooms and Toxins across the Freshwater-to-marine Continuum." *Integrated Environmental Assessment and Management*, vol. 19, no. 3, 2023, pp. 586-604. https://doi.org/10.1002/ieam.4651
9. Khan, Rabia M., *et al.* "A Meta-Analysis on Harmful Algal Bloom (HAB) Detection and Monitoring: A Remote Sensing Perspective." *Remote Sensing*, vol. 13, no. 21, 2021, article 4347. https://doi.org/10.3390/rs13214347
10. Risco-Martín, J. L., Esteban, S., Chacón, J., Carazo-Barbero, G., Besada-Portas, E., & López-Orozco, J. A. (2023). Simulation-driven engineering for the management of harmful algal and cyanobacterial blooms. *SIMULATION*, *99*(10), 1041–1055. https://doi.org/10.1177/00375497231184246
11. Randolph, Kaylan, *et al.* "Hyperspectral Remote Sensing of Cyanobacteria in Turbid Productive Water Using Optically Active Pigments, Chlorophyll a and Phycocyanin." *Remote Sensing of Environment*, vol. 112, no. 11, Aug. 2008, pp. 4009–19, https://doi.org/10.1016/j.rse.2008.06.002
12. Pyo, JongCheol, *et al.* "Cyanobacteria Cell Prediction Using Interpretable Deep Learning Model with Observed, Numerical, and Sensing Data Assemblage." *Water Research*, vol. 203, 2021, article 117483. https://doi.org/10.1016/j.watres.2021.117483
13. Ananias, Pedro H. M., *et al.* "ABF: A Data-driven Approach for Algal Bloom Forecasting Using Machine Intelligence and Remotely Sensed Data Series." *Software Impacts*, vol. 17, 2023, p. 100518, https://doi.org/10.1016/j.simpa.2023.100518
14. Zheng, Lei, Bo Hu, and Aizhong Ding. "A multi-factor data-driven prediction model for cyanobacteria blooms in lakes and reservoirs." *Desalin. Water Treat,* vol.189, June 2020, pp 207-216. https://doi.org/10.5004/dwt.2020.25621
15. National Water Quality Monitoring Council, 2023, Water Quality Portal, https://doi.org/10.5066/P9QRKUVJ
16. United States Environmental Protection Agency, "Recommended Human Health Recreational Ambient Water Quality Criteria or Swimming Advisories for Microcystins and Cylindrospermopsin." *United States Environmental Protection Agency*, May 2019, www.epa.gov/sites/default/files/2019-05/documents/hh-rec-criteria-habs-document-2019.pdf.
17. Pang, Chengfang, *et al*. "Multi-criteria Decision Analysis Applied to Harmful Algal Bloom Management: A Case Study." *Integrated Environmental Assessment and Management*, vol. 13, no. 4, 2017, pp. 631-639. https://doi.org/10.1002/ieam.1882
18. Ralston, D.K., and S.K. Moore. "Modeling Harmful Algal Blooms in a Changing Climate." *Harmful Algae*, vol. 91, Jan. 2020, article 101729. https://doi.org/10.1016/j.hal.2019.101729
19. Glibert, P. M. "Harmful Algae at the Complex Nexus of Eutrophication and Climate Change." *Harmful Algae*, vol. 91, 2020, article 101583. https://doi.org/10.1016/j.hal.2019.03.001.
20. Paerl, H. W., & Paul, V. J. Climate change: Links to global expansion of harmful cyanobacteria. *Water Research*, vol. 46, no. 5, Apr. 2012, pp.1349–1363. https://doi.org/10.1016/j.watres.2011.08.002
21. AryamanDShukla. "GitHub - AryamanDShukla/EProACH_24." *GitHub*, github.com/AryamanDShukla/EProACH_24.
22. "Eyes on the Bay: Harmful Algal Blooms Interactive Data Map." *Eyes on the Bay*, eyesonthebay.dnr.maryland.gov/eyesonthebay/habs.cfm. Accessed 20 Sept. 2023.
23. NCDEQ- Division of Water Resources. "Fish Kill & Algal Bloom Report Dashboard [WebMap]". Scale not Given. "North Carolina Department of Environment and Natural Resources." August 2023. ncdenr.maps.arcgis.com/apps/dashboards/7543be4dc8194e6e9c215079d976e7
24. *NOAA NCCOS Algal Bloom Beta/Experimental Products*. coastwatch.noaa.gov/cw_html/NCCOS.html?
25. "Algal Bloom Surveillance Map - Waterborne Hazards Control." *Waterborne Hazards Control*, 8 May 2024, www.vdh.virginia.gov/waterborne-hazards-control/algal-bloom-surveillance-map.