|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Analysis of Factors Influencing Harmful Algal Bloom Growth  Jason Curtis and Samuel Hughes / 12.08.2024  https://github.com/curtisjj42/mascoma\_lake |  |
|  |  |  |  |



# Table of Contents

[Table of Contents 2](#_Toc184576582)

[Introduction 3](#_Toc184576583)

[Methods 5](#_Toc184576584)

[Results & Discussion 8](#_Toc184576585)

[Conclusions & Limitations 10](#_Toc184576586)

[Appendix 11](#_Toc184576587)

|  |  |  |
| --- | --- | --- |
| Introduction |  | |
| Clean water access is a hallmark requirement for the survival and prosperity of communities as well as the ecological health of the region. Global trends in changing land-use to support rapidly growing populations show a significant decrease in natural wetland coverage, with expanding urbanization and agricultural needs demanding more and more land resources (Paludan et al. 2002; Zedler 2003; Hansson et al. 2005). The loss of natural wetlands has generally resulted in a minimization or loss of the central role wetlands play in protecting the ecological safety and stability of these clean water resources. Eutrophication, brownification, and climate change contribute additionally to the increasing amount of organic and inorganic compounds reaching inland bodies of water (Ekvall et al. 2013; Taranu et al. 2015; Kritzberg et al. 2020; Galloway and Cowling 2021). The presence of these compounds increases the nutrient availability in these systems and produces a significant increase in growth potential of phytoplankton species necessary to the structure of these local food webs. This increase in nutrient availability has led to a global rise in frequency of algal bloom events which reduces worldwide water quality and access to potable resources (Heisler et al. 2008; Li and Hong 2011).  Harmful Algal Blooms (HABs)  Algal blooms can arise from any of several species of phytoplankton – a type of photosynthetic prokaryote estimated to be responsible for nearly half of all global primary production and the primary base level of all aquatic systems (Field et al. 1998; Behrenfeld et al. 2001). Due to their ubiquity in aquatic ecosystems, observing and measuring algal growth is a consistent and dependable way to study the balance of aquatic systems (EPA, NOAA). Phytoplankton are responsible for using solar radiation to convert raw nutrients in the form of nitrogen (N) and phosphorus (P) into biomass to support all manner of animal life.  Algal blooms are often viewed as a nuisance, but larger occurrences (known as Harmful Algal Blooms or HABs) can result in catastrophic effects for their ecosystem. High concentrations of algae reduce the amount of sunlight reaching the water and the mass die-offs and decomposition of algae reduces the available dissolved oxygen (DO) in the water (US EPA, 2024). Often large algal blooms are followed by mass die-offs of fish and aquatic plants. Additionally, some species of phytoplankton known as cyanobacteria produce cyanotoxins as a growth byproduct. When ingested, these bacteria can lead to gastrointestinal damage, liver damage, neurological damage, and death (US CDC, 2024). For these reasons algal blooms and specifically cyanobacteria bloom frequency and severity are often considered to be strong indicators of hydrological system health and balance.  Causes and Related Factors  As mentioned above, one of the most widely recognized causes of HABs is nutrient flooding. A sudden influx of N and P may be delivered to the system by runoff containing fertilizer or more directly by industrial pollution. Other factors necessary for bloom occurrence are sunlight and warm temperatures. Blooms are often observed after large storms or other extreme weather events which cause mixing in the water column and disturb trapped nutrients in the lakebed. The interplay between the various natural and human-made factors is not well understood in today’s literature and discovering the relationship between these is critical to combatting rising numbers of HABs effectively to maintain and improve access to clean water.  The Role of Wetlands  One of the primary ways to control the occurrence of HABs involves nutrient restriction and filtering. Concentration of hydrophytic plants in wetland areas allows plants and microbes to absorb N and P directly from the water column and reduce the overall level flowing free in the water body. Bacteria in the wetland system complete the denitrification process by converting nitrogen stored as nitrate into nitrogen gas to be released into the atmosphere, limiting the available nitrogen and minimizing the growth potential of algae. Additionally, wetland vegetation provides competition for sunlight and other necessary nutrients, limiting their overall growth potential and reducing the likelihood of a bloom occurring. Because of the key role played by these systems in mitigating the potential for HABs, we posit that the loss of natural wetland resources is a direct causal factor in the rising frequency of harmful algal blooms. The following hypotheses were developed to investigate:   1. Cyanobacteria bloom activity has increased over the past two decades, with changes more pronounced in water bodies surrounded by lower wetland coverage. 2. Increased percentages of wetlands within a 1 km buffer are associated with more frequent and severe blooms. | |  |

|  |
| --- |
| Methods |
| The primary focus of this investigation is to study data gathered from real-world blooms in uncontrolled environments to better understand which factors commonly associated with HABs play the most significant role in their appearance. A secondary objective for this study is to understand the necessity of wetland conservation and construction as a strategy for mitigation of HABs by determining the relationship between loss of wetlands to the increasing incidence rate of HABs. The state of New Hampshire was selected for this study due to ongoing work by local governments to understand how to take action to preserve the health of their local watersheds.  Data Gathering  Initial data was provided by the New Hampshire Department of Environmental Services which maintains a database of all cyanobacteria blooms starting in the year 2003. This set contains date stamped advisories issued by the state indicating HAB activity in a body of water. Also included in this set are the measured algal concentration, location and lake information, duration of advisory, level of impact (warning of reduced quality or alert of actively toxic bloom), and species of bacteria responsible. This dataset contains 696 reported bloom events since 2003 primarily collected by municipal environmental monitoring organizations.  Other supporting datasets were gathered to help with analysis of the bloom data. Land cover/land use data for the years 1962, 1974, 1998, 2015, and 2021 in Strafford and Rockingham counties were collected from the NH GRANIT Clearinghouse and included GeoJSON and Geopackage formats of interpreted raster data. Finally, a dataset describing water and wetland resources in New Hampshire was collected from the National Wetlands Inventory (NWI) maintained by the U.S. Fish and Wildlife Service.  Data Preprocessing  The HAB occurrence data required significant preprocessing to be visualized and understood. Several columns are incomplete, with 26% of records missing for advisory end dates and advisory duration. Additionally, many of the bodies of water were named and designated incorrectly. For supporting datasets, preprocessing mostly consisted of clipping various datasets into target regions. The NWI dataset was used to support geocoding of the blooms. Preliminary EDA allowed early visualizations of bloom incidence trends to the present day.  The dataset used for both regression and classification modeling was prepared by amplifying the HAB dataset using the Cowardin System for Classification of Wetlands and Deepwater Habitats of the United States (Cowardin 1979). Each lake with a bloom was identified and geocoded using the address information provided by the NH DES dataset. A 1-kilometer buffer zone consistent with our hypothesized influential area was built around each lake and then spatially joined to produce a list of every wetland type in the lake system. We then one-hot encoded this feature list to generate a list of wetland types and structures in each hydrological system to generate our feature set describing each system.  Data Processing and Modeling  Following the initial cleaning, organization, and visualization, statistical and machine learning techniques were applied to investigate our hypotheses. Initial hypothesis testing involved splitting the dataset into two equivalent time periods to determine whether rising bloom rates are statistically significant. After quantifying the number of blooms in the state in the first 10-year period versus the second, we selected a Mann-Whitney U Test to determine statistical significance of the difference. This method was appropriate for several reasons: small set size, assumed non-normal data distribution with high variance, and continuity of target variable.  Next, we focused on quantifying changes in land use within the state. Due to data availability, we limited analysis to Strafford and Rockingham counties for consistency. This analysis involved calculating changes in land cover/land use from 1962-1998, using linear regression to model the speed of impervious surfaces in this period, and applying the model to data from 2015-2021. This was necessary because between 1998 and 2015, the state government overhauled its data collection and interpretation resulting in values an order of magnitude lower for the modern period. Assumptions made include linearity of the temporal trend in impervious surface coverage. The scaling factor was calculated by averaging the differences in prediction versus actual observed values. By applying this scaling to our data, we generated adjusted rates of change for the total period in the context of the historical data. We then assumed a linear relationship exists between creation of impervious surfaces and loss of wetland space so we could produce estimated values for rates of change in wetland resources in the state given unavailability of data between the years of 1998 and 2015.  For analysis of the bloom dataset itself, we applied multiple clustering and supervised modeling techniques. We evaluated clustering methods including K-means and DBSCAN algorithms to find meaningful clusters from the bloom occurrences. Inertial and silhouette scores were calculated to determine optimal numbers of clusters and PCA dimensionality reduction allowed us to further explore how various features influenced each data points’ position in its cluster. Supervised modeling techniques used include linear, gradient boosting, and support vector machine modeling. Regression modeling was completed with the purpose of testing the robustness of our dataset as well as the predictive capabilities. Performance was evaluated using R squared calculations, MAE, and RMSE. Feature importance for our best performing models was analyzed by evaluating feature weighting and SHAP values. At this stage of research, outliers were not removed from the dataset due to the already small size and risk of losing critical insights.  Classification modeling was completed using two methods to evaluate how well we could classify HAB risk in lakes based on wetland features contained within each lake system. Four risk classes were created using two differing methods to account for bias in clustering results. Method 1 involved calculating the mean HAB incidence for each cluster and assigning a risk class to each cluster based on mean bloom count. Method 2 involved manually dividing the dataset into bins using the mean bloom count as a centerpoint for the bin parameters. In theory, this would produce a relatively normal class distribution and help verify or dismiss earlier assumptions that the dataset is non-normally distributed. By comparing these methods of lake classification and then training classification models we were able to evaluate how well the clusters describe risk similarities in the lake systems. | |

|  |
| --- |
| Results & Discussion |
| Initial statistical analysis of the overall set revealed that not only has there been a meaningful increase in bloom activity but the increase itself is large. The Mann-Whitney U test demonstrated the statistical significance of the period analysis with a U statistic of 11 and a p-value of 0.0022. This allows rejection of the null hypothesis that there is no increase in HABs. With a rank-biserial correlation of -0.8, we can assert that this association is strongly skewed towards period two.  Our results of the variation in impervious surface development and wetlands loss showed increases across the board in both the rate of impervious surface creation and wetland loss. The degree of change was quite different between the two townships of interest, with Strafford showing nearly a tenfold increase in impervious surface coverage while Rockingham’s degree of change was closer to double. This is likely due to Rockingham being made up of more urban spaces than Strafford and already having been concentrated with impervious cover prior to data collection start. The degree of change brings Strafford more closely in line with the land distribution in Rockingham which indicates an overall expansion of urbanized regions in New Hampshire.  Figure 1: Comparison of HABs between 10-year periods  After initially trialing both K-means and DBSCAN algorithms for clustering the bloom dataset, K-means was determined to be the preferred method. DBSCAN produced one dense cluster and classified everything else as outliers. While informative for understanding the data distribution, it provided next to no information about the blooms themselves and was discarded. Based on inertial decay and silhouette scores, we determined 4 clusters to be the best solution. Our clusters themselves were not well defined, but our analysis of the principal components defined during dimensionality reduction demonstrated two primary insights. For the first component, the most impactful feature was wetland coverage. This was confirmed by our regression modeling results which identified wetland concentration as the highest impact feature for determining bloom incidences. The second component was largely driven by wetland classifications with the most impactful features being systems containing permanently flooded wetland systems with unconsolidated bottoms.  Figure 2: K-means 4 cluster solution  Regression modeling proved to be largely a failure on this dataset. Due to the small size and high influence of outlier lakes such as Lake Winnipesauke - which made up a large portion of the set – our models were too volatile to reliably implement. Because outlier removal would constitute nearly 1/3rd of the rows in our set, we retained these outliers in our set so we could investigate more closely why the models behaved this way. Surprisingly, we found that not only was wetland coverage highly influential to the predictive capabilities of the model it was in the wrong direction: more wetlands correlated highly with more HABs. By evaluating SHAP values for each feature we discovered that this was largely driven by two of our outlier systems as we suspected, though no pattern was discerned in the feature elsewise. Another feature of note was wetland ratio: the ratio of wetlands to lake body in the system. We found that a high wetland ratio was also highly correlated with bloom incidence. This context allowed us to better understand that the raw area of wetlands was less impactful than initially thought. High wetland ratio inversely correlates with lake size, indicating that smaller lakes are less resilient to blooms.  Finally, classification modeling was significantly more successful at generating useful and robust solutions with fewer overfitting problems. As discussed in our methods section, we explored two methods for class creation with classes based on the clusters much the more successful of the two. Standard gradient boosting provided the most consistent results, performed well across all evaluation metrics, and additionally was the only method which consistently produced high accuracy scores in all 4 classes. By contrast, the manual creation method proved highly ineffective and indicates more complex relationships happening between the features than we predicted. This is likely due to the breadth of the feature set. Given this level of success it would be informative to expand our model to datasets from other New England states to see if the success holds or if our method was too biased towards our training data. |  |

|  |  |  |
| --- | --- | --- |
| Conclusions & Limitations |  | |
| With regards to our research hypotheses, success was mixed. Our first hypothesis held up well under scrutiny, but our second hypothesis proved to be more complex than the scope of this project allowed. We have confirmed that regional instances of HABs are statistically up from previous decades but were unable to identify the specific role that wetlands play in these systems at a large scale. This was a direct result of an overly simplistic hypothesis which address a clearly more complex interaction than we had previously understood.  Overall, this project struggled from several limitations that prevented higher quality analysis. Our study involved a high degree of speculation, limited data, and a few assumptions and decisions made from necessity due to the scarcity of data. As discussed previously, we did not move ahead with outlier removal during our modeling because of the destructive effect it had on our dataset. However removal of these lakes would have been a key step in improving our actual insights and information extraction from our models. Since these lakes drove the model behavior, it is entirely possible that not accounting for them deeply affected the feature contributions to our results from all models.  Re-generating the cluster solutions without outlier lakes would be an ideal next step to recontextualize the analysis done in this report. This would allow us to further evaluate the success of our cluster-based classification system and provide further insights into the viability of that analysis. Bloom risk classification was not a primary project objective but based on data availability it became by far the most successful portion of this study and has the greatest potential for real-world use. Further research into classifying HAB risk is a promising project and should be explored by augmenting these data with similar metrics from geographically differentiated regions to build a more robust understanding of how our classes are structured and how to address the growing problem of HABs.  This project continues to be an ongoing effort to understand the rising rates of HABs across the globe. While much of our analysis was unsuccessful at this stage it has directly informed the type and quality of data necessary for improving the potential of further investigation. Appendix | |  |

Bibliography

Behrenfeld MJ, Randerson JT, McClain CR, Feldman GC, Los SO, Tucker CJ, Falkowski PG, Field CB, Frouin R, Esaias WE, Kolber DD, Pollack NH (2001) Biospheric primary production during an ENSO transition. Science 291:2594–2597

CDC. (2024, May 15). *Harmful Algal Blooms and Your Health*. Harmful Algal Bloom (HAB)-Associated Illness. <https://www.cdc.gov/harmful-algal-blooms/about/index.html>

Cowardin, L.M., Carter, V., Golet, F.C., and LaRoe, E.T., 1979. Classification of Wetlands and Deepwater Habitats of the United States

Ekvall MK, Martin JD, Faassen EJ, Gustafsson D, Lurling M, Hansson LA (2013) Synergistic and species-specific effects of climate change and water colour on cyanobacterial toxicity and bloom formation. Freshwater Biology 58:2414–2422

EPA. (2024, January 3). *The Effects: Dead Zones and Harmful Algal Blooms*. US EPA. <https://www.epa.gov/nutrientpollution/effects-dead-zones-and-harmful-algal-blooms>

Field CB, Behrenfeld MJ, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237–240

Galloway JN, Cowling EB (2021) Reflections on 200 years of Nitrogen, 20 years later this article belongs to Ambio’s 50th anniversary Collection. Theme: Eutrophication Ambio 50:745–749

Heisler J, Glibert PM, Burkholder JM, Anderson DM, Cochlan W, Dennison WC, Dortch Q, Gobler CJ, Heil CA, Humphries E, Lewitus A, Magnien R, Marshall HG, Sellner K, Stockwell DA, Stoecker DK, Suddleson M (2008) Eutrophication and harmful algal blooms: a scientific consensus. Harmful Algae 8:3–13

Kritzberg ES, Hasselquist EM, Skerlep M, Lofgren S, Olsson O, Stadmark J, Valinia S, Hansson LA, Laudon H (2020) Browning of freshwaters: consequences to ecosystem services, underlying drivers, and potential mitigation measures. Ambio 49:375–390

NOAA. (2016, April 27). *What is a harmful algal bloom?* Noaa.gov; noaa. <https://www.noaa.gov/what-is-harmful-algal-bloom>

Li Y, Hong HS (2011) Nutrient sources and composition of recent algal blooms and eutrophication in the northern Jiulong River, Southeast China. Marine Pollution Bulletin 63:249–254

Paludan C, Alexeyev FE, Drews H, Fleischer S, Fuglsang A, Kindt T, Kowalski P, Moos M, Radlowki A, Stromfors G, Westberg V, Wolter K (2002) Wetland management to reduce Baltic sea eutrophication. Water Science and Technology 45:87–94

Taranu ZE, Gregory-Eaves I, Leavitt PR, Bunting L, Buchaca T, Catalan J, Domaizon I, Guilizzoni P, Lami A, McGowan S, Moorhouse H, Morabito G, Pick FR, Stevenson MA, Thompson PL, Vinebrooke RD (2015) Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. Ecology Letters 18:375–384

Zedler JB (2003) Wetlands at your service: reducing impacts of agriculture at the watershed scale. Frontiers in Ecology and the Environment 1:65–72

‌

Graphs and Figures

A graph with a line

Description automatically generatedA graph with red lines

Description automatically generated

Figure 3: Inertial decay for K-means

Figure 4: Cluster silhouette scores

A graph with blue and white bars

Description automatically generated

A graph with blue and white lines

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| County | Period | Impervious Change (m²) | Rate (m²/year) | Wetlands Change (m²) | Rate (m²/year) |
| Strafford | 1962-1998 | 130,000,000 | 3,610,000 | -68,700,000 | -1,900,000 |
|  | 2015-2021 | 4,800,000 | 800,000 |  |  |
|  | 1962-2021 | **1,033,060,540\*** | **17,509,500\*** | *-594,047,150\*\** | *-10,068,595\** |
| Rockingham | 1962-1998 | 407,000,000 | 11,300,000 | -320,000,000 | -8,800,000 |
|  | 2015-2021 | 6,400,000 | 1,070,000 |  |  |
|  | 1962-2021 | **1,218,031,200\*** | **20,644,596\*** | *-700,411,591\*\** | *-11,871,382\*\** |

Table of change rates for Impervious surface and wetlands coverage in Strafford and Rockingham counties

*\*values are adjusted to account for overhaul of state repository management*

**\*\*values are estimated from prior data (actual data unretrievable)**

A graph with blue dots and red line

Description automatically generated

A graph with a line graph and numbers

Description automatically generated with medium confidence