University of Guelph

Establishing Potential Cyanobacterial Bloom Triggers in Lake Erie using Spatial and Regression Analysis

Connor Campbell
Floyd Pinto
Gurjap Singh

Supervisor: Dr. Ben DeVries

Winter-2020

Table of Contents

Abstract

Cyanobacterial algal blooms in Lake Erie severely impact the economy of various industries in the surrounding areas as well as human and wildlife health. The tourism industry alone faces an annual loss of up to $110 million (Smith et al., 2019). Tourism decreases when the beaches become unsuitable for swimming and even more when fishing prospects decrease after the waters become uninhabitable for sensitive fish species (Smith et al., 2019). Moreover, the Microcystin toxin produced by the blooms can poison marine life, humans, and human resources. It would be greatly beneficial for both tourism revenue and human health to be able to estimate cyanobacterial bloom quantity and density using data collected from environmental variables to help decision makers design strategies that can reduce the severity of Cyanobacteria blooms and preserve the economy of the industries that depend on lake erie. The aim of this study was to first identify the environmental variables that contribute to Cyanobacteria blooms and investigate how their quantities impact the quantity and density of the blooms. This study implements a multiple linear regression analysis to identify the relationships between Cyanobacteria density derived from satellite imagery and environmental factors such as phosphorus, nitrogen, and water temperature. Results of the regression analysis show a positive linear relationship between nutrient concentrations of phosphorus and nitrogen and Cyanobacteria density. It was also seen that spatial displacement of the blooms due to water current and wind displacement (CBC, 2019) tended to undermine the positive linear relationship between nitrogen, phosphorus, and Cyanobacteria density. A geographically weighted regression that was additionally run, made evident that phosphorus had a higher weighting in the model the closer it was to the Maumee Bay, indicating that the Maumee River inlet is a likely cause for high phosphorus levels in run-off during bloom season.

1. Problem Context

In recent decades, climate change has become a globally growing concern, especially for water bodies. Increased water surface temperature, decreased water pH, and increased frequency of algal blooms are all consequences of a rise in global temperatures (Moore et al., 2008; Parry et al., 2007). The increase in algal blooms occur when changes in temperature alter the water column’s stratification causing mixing between the nutrient rich bottom layers of water with warmer upper layers of water. This mixing then supplies Cyanobacteria located in the warmer upper layers of water with nutrients required for growth, thus triggering bloom events (Moore et al., 2008).

Cyanobacteria, a photosynthetic organism (Vermaas, 2001), that usually thrives in warm, shallow, calm waters like beaches, bays, and coasts (Chen et al., 2014; Bennington-Castro, 2015). Climate change increases the frequency and severity of blooms by warming waters, increasing available solar energy, increasing CO2 as inputs for photosynthesis, and increasing potential blooming areas when rising global water levels expand bays (Environmental Protection Agency, 2019). Cyanobacteria blooms in Lake Erie is a severe problem that has led to beach shutdowns, loss of drinking water for some cities, and loss of an estimated annual revenue of $272 million (Smith et al., 2015; Haggert, 2019). In severe cases, Cyanobacteria can bloom into a Harmful Algal Bloom (HAB). HABs are more detrimental than their normal counterparts as they may produce neurotoxins, liver toxins, and cell toxins, that can affect drinking water, enter the food chain through fish, kill large fish populations, and poison marine ecological systems (Sivonen & Jones, 1999; Foster, 2013). Moreover, Cyanobacteria blooms can also create dead zones which are hypoxic environments caused by bloom overgrowth; this can restrict other organisms' access to O2 and sunlight (Altieri & Gedan, 2014).

Even though a multitude of factors have been linked to Cyanobacteria blooms, such as Phosphorus, Nitrogen, water temperature, water depth, and time of year; there are still knowledge gaps in this field of study. One of these knowledge gaps include the factors that cause the Microcystin producing Cyanobacteria genus (Microcystis) to prevail in competition amongst other Cyanobacteria genus during a bloom event. It is important to fill this knowledge cap because the Microcystin toxin is one of the most devastating byproducts of a harmful algal bloom, and is thus a point of interest (Harke et al., 2016).

2. Research Purpose

To investigate the relationship between environmental variables and the density of cyanobacteria through the use of a GIS Model.

3. Research Objectives

Objective 1)

Identifying factors and variables that impact the density of algal blooms.

Objective 2)

Developing a GIS based model that can be used to evaluate the density of algal blooms in the western basin of Lake Erie.

Objective 3)

Apply the model to determine whether the observed variables impact the behaviour of algal bloom density.

Objective 4)

Evaluate the Strengths and Weaknesses of this study.

4. Study Area

Located on the international border between Canada and the United States is one of the Great Lakes, Lake Erie. It is the shallowest of all 5 of the Great Lakes and its biggest inlet is the Detroit River which feeds into its western basin. Upon entering, the water flows east until it reaches its biggest outlet in the east basin, the Niagara River (Watson et al, 2016).

In the 1960s, Lake Erie was heavily affected by algal blooms due to excessive amounts of phosphorus pollution, and the toxins produced by the algal blooms caused the inhabiting fish to die in large numbers (Foster, 2013).

In 2011 nearly 20% of Lake Erie was covered in a type of Cyanobacteria known as *Microcystis*, which secretes a toxin that can cause extreme sickness (Foster, 2013). The Maumee, Sandusky, and Raisin River inlets are significant contributors to agricultural runoff for the western basin of Lake Erie, which will be the prime focus of this study (Bosch et al., 2014).

Lake Erie's western basin, shown in Figure 1, has experienced a large amount of worsening algal blooms (Foster, 2013). The recent increase in algal blooms is large enough that it can be seen through satellite imagery, as shown in Figure 2. The most heavily impacted area in Lake Erie's western basin is the area near Maumee Bay which receives a large amount of agricultural nutrient runoff from the Maumee, Sandusky, and Raisin River inlets (Bosch et al., 2014). This area can be seen in Figure 2 and will be the primary focus of our research. Focusing on this area should provide both nutrient and bloom data to then create a model that will provide insight into the causes behind algal bloom.

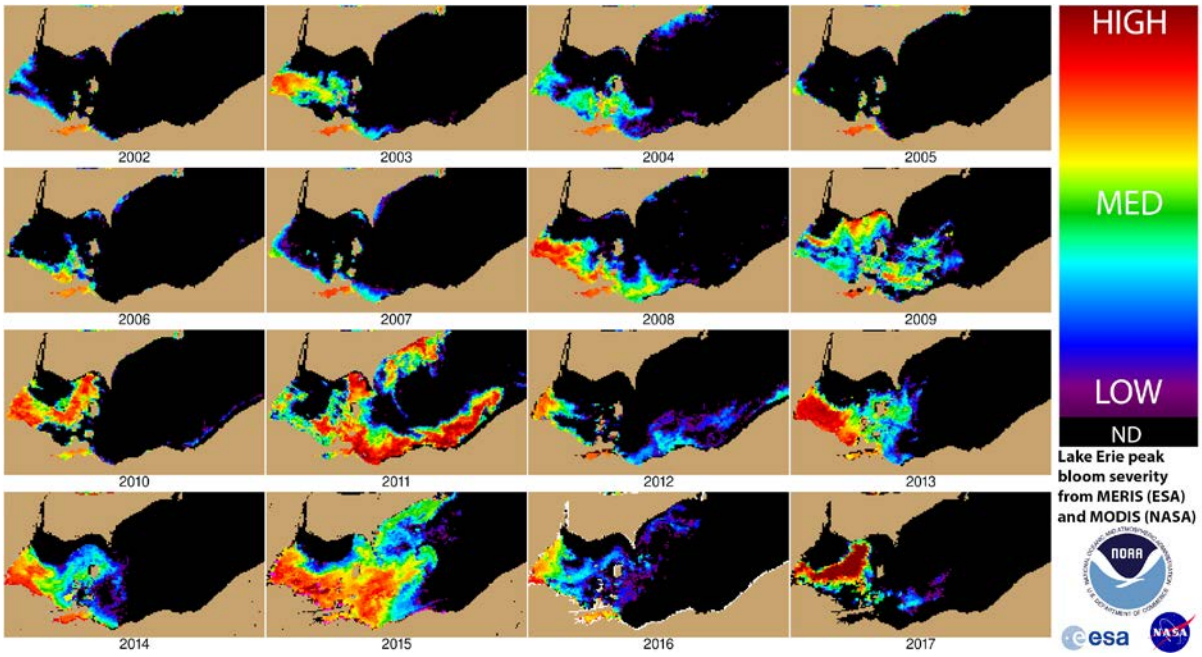

***Figure 1**: A time series of Lake Erie's bloom severity. Shown to be increasing from 2002 to 2017.*

Adopted from *https://news.osu.edu/noaa-partners-predict-large-summer-harmful-algal-bloom-for-western-lake-erie/*

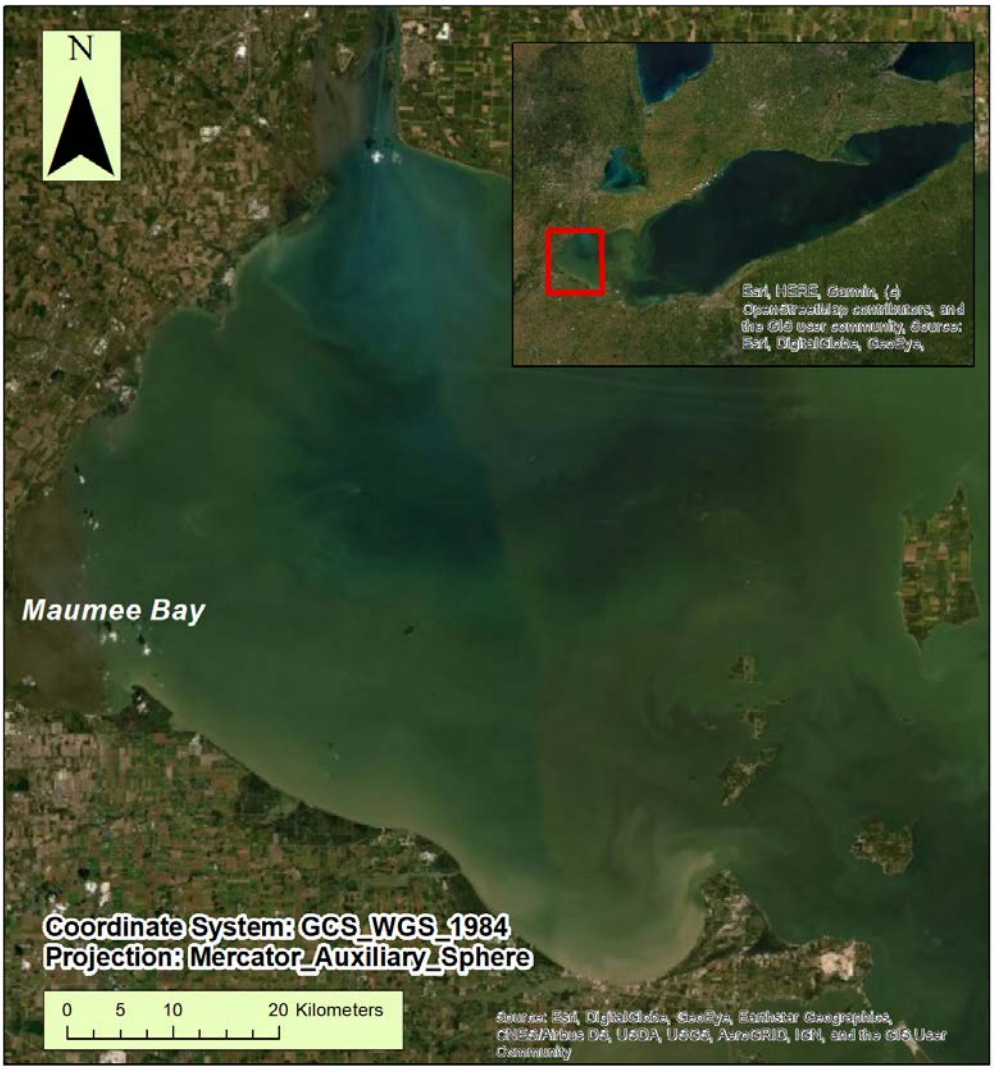

***Figure 2**: Imagery of the western most area of Lake Erie*

5. Research Approach

This study accomplished the research objectives using the following approaches:

Objective 1: Identify factors and variables that relate to the density of algal blooms.

Developing a regression model requires two key pieces of data: A response variable and a set of explanatory variables. The response variable of the will be the focus of the study. The study will attempt to explain why this phenomenon occurs, and it will do that by comparing the response variable to the explanatory variables. This study aims to establish a relationship between the behaviour of the explanatory variables in

relation to the response variable. The response variable for this study was the Cyanobacteria Index (CI), which was used as a proxy for the presence of algal blooms. The explanatory variables were identified using scientific literature; Based on our research, which is outlined below, we believed phosphorus, nitrogen, and water temperature would have a statistically significant relationship with algal bloom growth.

Response Variable -- Cyanobacteria Index

Cyanobacteria index was used as a proxy for algal blooms (Ogashawara, 2019). The data used is a product that indicates the density of Cyanobacteria present within each raster cell. Part of the model for this project reclassifies the imagery to find the maximum density value in a specified area.

Explanatory Variable -- Phosphorus

Phosphorus has been observed to be a likely accelerant in the growth of algal blooms (Bachmann et al., 1974). It has also been identified by Environment Canada to be the primary driver for eutrophication in Lake Erie causing Cyanobacteria blooms (Government of Canada, 2018).

Explanatory Variable -- Nitrogen

Like phosphorus, nitrogen is a known bio-element to play a role in the growth of HABs (Herrero et al., 2001), and nitrogen levels are known to have increased in Lake Erie (Lane, 2019).

Explanatory Variable -- Water Temperature

Water temperature is an indicator of aquatic seasonal turnover which is when most problematic HABs occur (Pitcher et al., 2010), changes in temperature due to climate change impacts the timing and fullness of a complete turnover which can result in changes in nitrogen and phosphorus ratios as well as increased water temperatures for longer periods, both of which are favourable conditions for Cyanobacteria growth (Posch et al., 2012; Climate Change and Harmful Algal Blooms, 2019).

***Table 1**: Data layers to be used in the model*

Dataset	Type	Source	Publisher	Spatial Resolution	Temporal Resolution	Year
Cyanobacteria Index	GeoTIFF	Copernicus Sentinel-3a & 3b satellites, OLCI sensor	NOAA	300m	Daily	May to October 2017 to 2019
Phosphorus	CSV	NOAA WEXX sampling stations	NOAA	0.01 µg P/L	Weekly	May to October 2012 to 2018
Nitrogen	CSV	NOAA WEXX sampling stations	NOAA	0.01 mg N/L	Weekly	May to October 2012 to 2018
Water Temperature	CSV	NOAA WEXX sampling stations	NOAA	0.01°C	Weekly	May to October 2012 to 2018

The first dataset consists of GeoTIFF raster images for Western Lake Erie spanning 2017-2019. The images have been processed and reclassified by NOAA through their in-house Cyanobacteria Index algorithm (Vander Woude et al., 2019). The raster values representing the Cyanobacteria Index will be used in our model to observe the response variable, the average density of Cyanobacteria blooms.

The three explanatory variables of phosphorus, nitrogen, and water temperature will be obtained from nine National Oceanic and Atmospheric Administration (NOAA) nutrient sampling stations dispersed in and around Maumee Bay in Lake Erie's Western Basin. These points provide samples either weekly or biweekly. The dispersion of the nine NOAA water quality buoys is illustrated below in Figure 3.

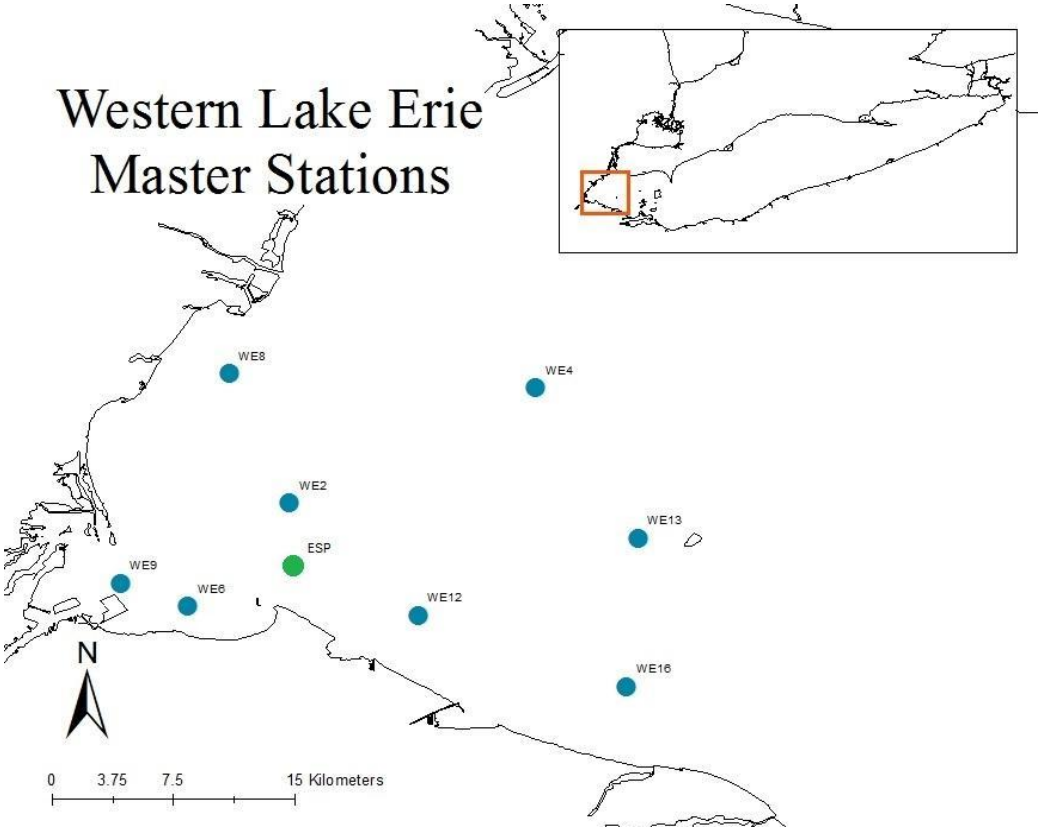

Figure 3: *A map created by NOAA illustrating the locations of their sampling stations.* Adopted from *https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/graphics/WLE2018Map.jpg*

Objective 2: Develop a GIS based model that can be used to evaluate the density of algal blooms in the western basin of Lake Erie.

The model for this study consisted of two main components. The first component is the spatial analysis while the second one is the statistical analysis.

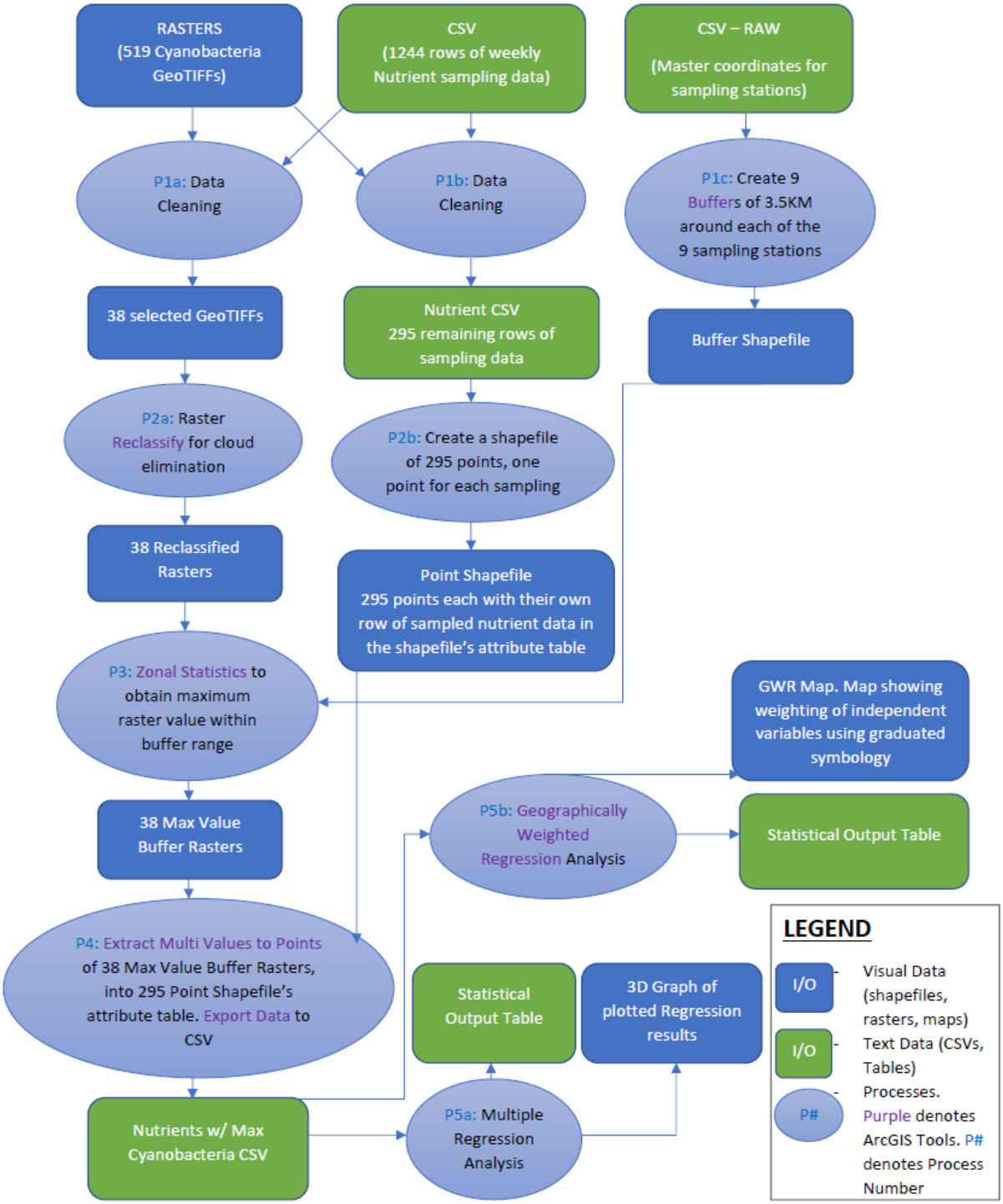

Figure 4: *A diagram visualizing our workflow model.*

The first step of the model, referred to as Process 1a and Process 1b in figure 4 was to sort and clean our data inputs. These inputs consisted of Cyanobacteria Index GeoTIFFs with daily temporal resolution during bloom season (May-October) of 2017-2019, and a CSV file with nutrients sampled weekly during bloom season of 2012-2018, with samplings occurring at 9 different stations. The first step in cleaning was to select the rasters which had corresponding dates with the weekly sampled data. The second

step was removing rasters with excessive cloud cover. After this step, 38 rasters remained which then led to the first step of cleaning the CSV. The CSV & rasters matched up temporally prior, but due to rasters being removed, corresponding dates within rows of the CSV also had to be removed. 295 rows remained after this clean-up. The CSV also had some inaccuracies with various sampling row's coordinates which needed to be manually replaced with the correct coordinates that were provided in a coordinate master file for the sampling stations.

The next part of the model consisted of spatial processing and analysis. Process 1c, as shown in figure 4, was to plot the coordinates of the 9 sampling stations and create a 3.5 km buffer around each of them in a single shapefile. The distance of 3.5 km was selected by using the distance between the 2 closest sampling sites, we wanted to maximize buffer size without having overlap with other sampling sites. Process 2a (refer to figure 4 & 5) was to reclassify the rasters to remove cloud cover and other invalid values so that these values would not interfere with statistical raster calculations later on. Process 3 was to use the Zonal Statistics tool in ArcMap to cut each of the 38 reclassified rasters into a raster containing nine 3.5 km buffers. Each buffer was assigned the raster's maximum value within said buffer's range. Maximum was selected as the statistic type to measure potential of bloom density because blooms are prone to wind/current displacement and there is already a low temporal resolution of 1 week giving bloom displacement a fairly good probability of occurring. Process 2b (refer to figure 4) created a shapefile of 295 points with an attribute table of nutrient sampling data from the CSV. Process 4 (figure 4) used the 'Extract Multi Values to Points' tool in ArcMap to append the values of the maximum buffers into the attribute table of the 295 nutrient points shapefile. These appended values represented the Maximum observed Cyanobacteria Index within a 3.5 km buffer range, and was exported to a CSV. This Nutrients with Max Cyanobacteria CSV however was still using the scaled down 8-bit raster values to represent Cyanobacteria, and so consequently had to be rescaled to true Cyanobacteria density values in cells/ml by using the re-scaling equation from the data provider NOAA. See equation (1):

$$True\ Value\ =\ 10^{(3\div 250\times(8-bit\ Value)-4.2)}\times 100000000 \quad (1)$$

This re-scaled nutrients with Cyanobacteria CSV was used in a multiple regression analysis(described in the next paragraph), as well as a geographically weighted regression analysis to produce a 3D graph that plots the relationship between the density of algal blooms and the explanatory variables.

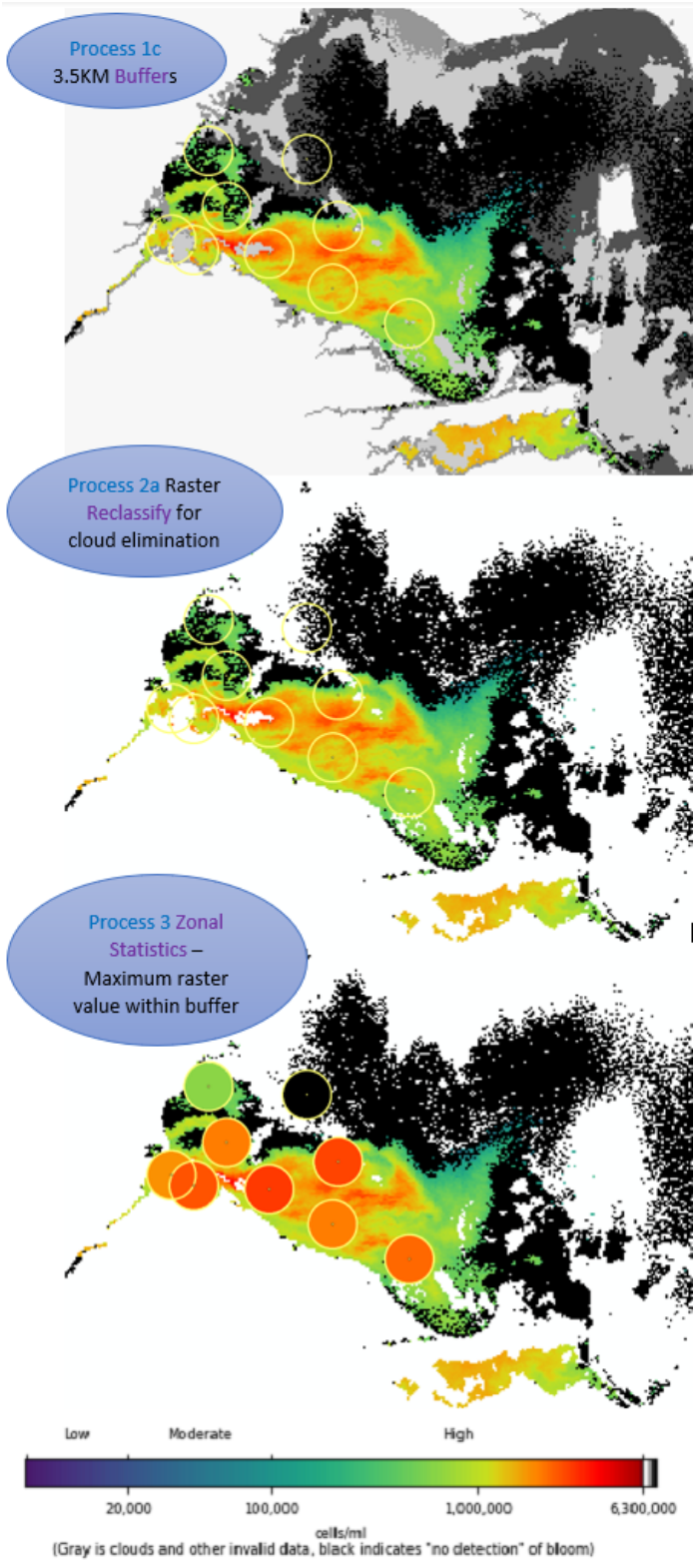

Figure 5: *Spatial processes, as referred to in Figure 4. Maximum raster value within buffer analysis. Raster value represents Cyanobacteria density in cell/ml as shown in legend.*

The second, more statistical component of our model was the multiple regression analysis. The regression model was chosen because of its suitability to finding the causal effect relationship between an explanatory variable and multiple different explanatory variables (Ray, 2019). Regression analysis also provides certain benefits like indicating the *significant relationships* and *strength of impact* of multiple explanatory variables on a response variable (Ray, 2019). In multiple linear regression there will be more than one explanatory (X) variable (Khan, 2012), as shown in equation (2):

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + u \quad (2)$$

where:
$\hat{Y}$ = predicted value of response variable (Cyanobacteria blooms)
a = the intercept
X_1 = first explanatory variable (water temperature)
X_2 = second explanatory variable (phosphorus)
X_3 = third explanatory variable (nitrogen)
u = the regression residual.

This equation calculates the line of best-fit for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to that line of best-fit (Ray, 2019). This specific kind of analysis, known as Ordinary Least Squares (OLS) was used to perform the multiple regression analysis. We also performed a second analysis using Geographically Weighted Regression (GWR). If the relationship between the response and explanatory variables is nonstationary, GWR will most likely result in a better model than OLS regression. GWR also results in a spatial visualization of our results, which is useful since the output of the OLS regression is solely statistical.

Objective 3: Apply the model to determine whether a correlating relationship is present between the observed variables.

After completing the previous objective, we had a model for processing our data and the means for performing both an OLS regression and a Geographically Weighted Regression.

OLS regression analysis identifies whether or not the predictors are meaningful additions to the model. OLS calculates a coefficient for each variable, and each variable has a probability associated with it. This probability, known as p-value, is a number between 0 and 1. Examining p-value helps draw conclusions in relation to the null hypothesis, which states that there is no relation between the variables being studied. The opposite of this is the alternative hypothesis, which states that the explanatory variable had a significant effect on the response variable. Any variable with a p-value below 0.05 is statistically significant. This means that there is less than a 5% probability that the null hypothesis is correct. If this is the case, then we can reject the null hypothesis and accept the alternative hypothesis (Mcleod, 2019). The analysis of the p-values of each explanatory variable can be used to determine whether or not a

statistically significant relationship exists between it and the response variable. If not, we can hypothesize as to why this discrepancy may have occurred and then adjust the model if necessary. The output of the regression also indicates whether or not the observed relationships likely vary over space. In this case, a geographically weighted regression would be performed to see if it has any impact on the p-values of the explanatory variables.

Unlike OLS, the output of a GWR provides a thematic visualization of the coefficients produced from the analysis. The coefficients produced from this analysis would allow us to visualize how each explanatory variable affected the response variable and how that relationship varied over space.

Objective 4: Evaluate the Strengths and Weaknesses of the model.

OLS automatically calculated a series of statistical checks in order to determine whether or not a regression model was properly specified and trustworthy. We examined these checks to confirm whether or not we created a reliable model.

1. Are the explanatory variables helping the model?

OLS calculates a coefficient probability (p-value) for each variable. Any value less than 0.05 was considered statistically significant (Mcleod, 2019). A Koenker test is also performed to determine if there is a nonstationary relationship in the data. If so, a GWR should be applied.

2. Do the relationships have positive or inverse correlation?

Examine the signs in front of the variable coefficients (positive or negative)

3. Is the model biased?

If the Jarque-Bera statistic, which determines whether or not residuals were normally distributed, are statistically significant, then the model is biased and considered untrustworthy. A biased model could potentially indicate that there are key explanatory variables missing from the model, or that not enough data is present.

4. Do we have all key explanatory variables?

The spatial autocorrelation tool is run to determine whether or not there is statistically significant spatial autocorrelation (clusters in the residuals). If clustering occurs, it is a symptom of misspecification, and misspecification indicates that key explanatory variables are missing.

5. How well are we explaining the response variable?

Adjusted R-Squared and Akaike's Information Criterion (AIC) was used to compare model performance. AIC is a relative value that only matters in comparison to other models with the same response variable. If changing a variable decreased AIC, it was considered a stronger model (Akaike, H. 1998). R^2 was used to see how much of the response variable was being explained by the response variable. R^2 is an indicator of the amount of variance explained by our model (Seber & Lee, 2012). A R^2 value like 0.2 means the model only accounts for 20% of the response variable variance. That value may seem low, but that information is still useful. It tells us that there are other sources of variance that are unaccounted for, which is likely to happen when examining a problem as complex as algal blooms. Finally there are residuals, which are the difference between the predicted values and the actual values. They can be thought of as an error margin. The sum of all residuals will be averaged, and defined as a percentage to define this margin. A margin of roughly 5% average residual error will be considered a model containing significant relationships, as a confidence level of 95% is widely used (Vijalapuram, 2019).

6. Research Findings

In our research we decided to use multiple regression models, each with different explanatory variables and temporal resolutions, in order to determine which yielded the most desirable p-values and Adjusted R-Squared. The first regression used the explanatory variables phosphorus, nitrogen, and water surface temperature for the years 2016, 2017 and 2018. The results are provided in Table 2.

Table 2: *Regression results from using phosphorus, nitrogen, and water surface temperature as variables for the years 2016, 2017 and 2018*

Variable	**P-value**	**Adjusted R-Squared**	**Average Residual Error**
Total Phosphorus (µg P/L)	0.003011033*	0.137594032	12.7%
Particulate Organic Nitrogen (mg/L)	0.00153796*		
CTD Temperature (°C)	0.726741148		

As shown in table 2, the first regression ended up having a relatively low Adjusted R-Squared at about 0.13. Adjusted R-Squared is more reliable than regular R-squared because it has adjusted the statistic based on the number of explanatory variables in the model. More importantly, the p-value for temperature was well above the allowable threshold of 0.05. In order to account for this, we removed temperature and ran a second model with only phosphorus and nitrogen as the explanatory variables.

Table 3: *Regression results from using Phosphorus and nitrogen as variables for the years 2016, 2017 and 2018*

Variable	**P-value**	**Adjusted R-Squared**	**Average Residual Error**
Total Phosphorus (µg P/L)	0.000402975*	0.142848688	12.7%
Particulate Organic Nitrogen (mg/L)	0.002427994*		

As shown in table 3, all the included variables now have acceptable p-values, but the Adjusted R-Squared is still low. An Adjusted R-Squared value of 0.14 means that we are only explaining about 14% of the response variable with our explanatory variables. This shows that we are at least explaining some of the variance with our model, though 14% is still a relatively low percent. We wanted to see if we could improve the R-Squared value while still maintaining significant p-values, so we decided

to run three different models. There would be one model for each year of data: 2016, 2017, and 2018.

***Table 4**: Regression results separated by year using phosphorus and nitrogen as variables for the years 2016, 2017 and 2018*

2016 - Phosphorus and Nitrogen			
Variable	**P-value**	**Adjusted R-Squared**	**Average Residual Error**
Total Phosphorus (µg P/L)	0.000402975*	0.675972842	6.1%
Particulate Organic Nitrogen (mg/L)	0.002427994*		
Regression Equation: Cyanobacteria(cells/ml) = 8307(Phosphorus mg/L) + 1660394(Nitrogen µg P/L) - 372147			
2017 - Phosphorus and Nitrogen			
Variable	**P-value**	**Adjusted R-Squared**	**Average Residual Error**
Total Phosphorus (µg P/L)	0.675108567	0.016939978	18.9%
Particulate Organic Nitrogen (mg/L)	0.112435445		
Regression Equation: Cyanobacteria(cells/ml) = 1628(Phosphorus mg/L) + 412094(Nitrogen µg P/L) + 1406113			
2018 - Phosphorus and Nitrogen			
Variable	**P-value**	**Adjusted R-Squared**	**Average Residual Error**
Total Phosphorus (µg P/L)	0.589785256	0.424457641	2.6%
Particulate Organic Nitrogen (mg/L)	1.45216E-05		
Regression Equation: Cyanobacteria(cells/ml) = -525(Phosphorus mg/L) + 1509188(Nitrogen µg P/L) + 55180			

As shown in table 4, for each of the three years that we have data for, only 2016 has produced a high Adjusted R-Squared value with acceptably low p-values.

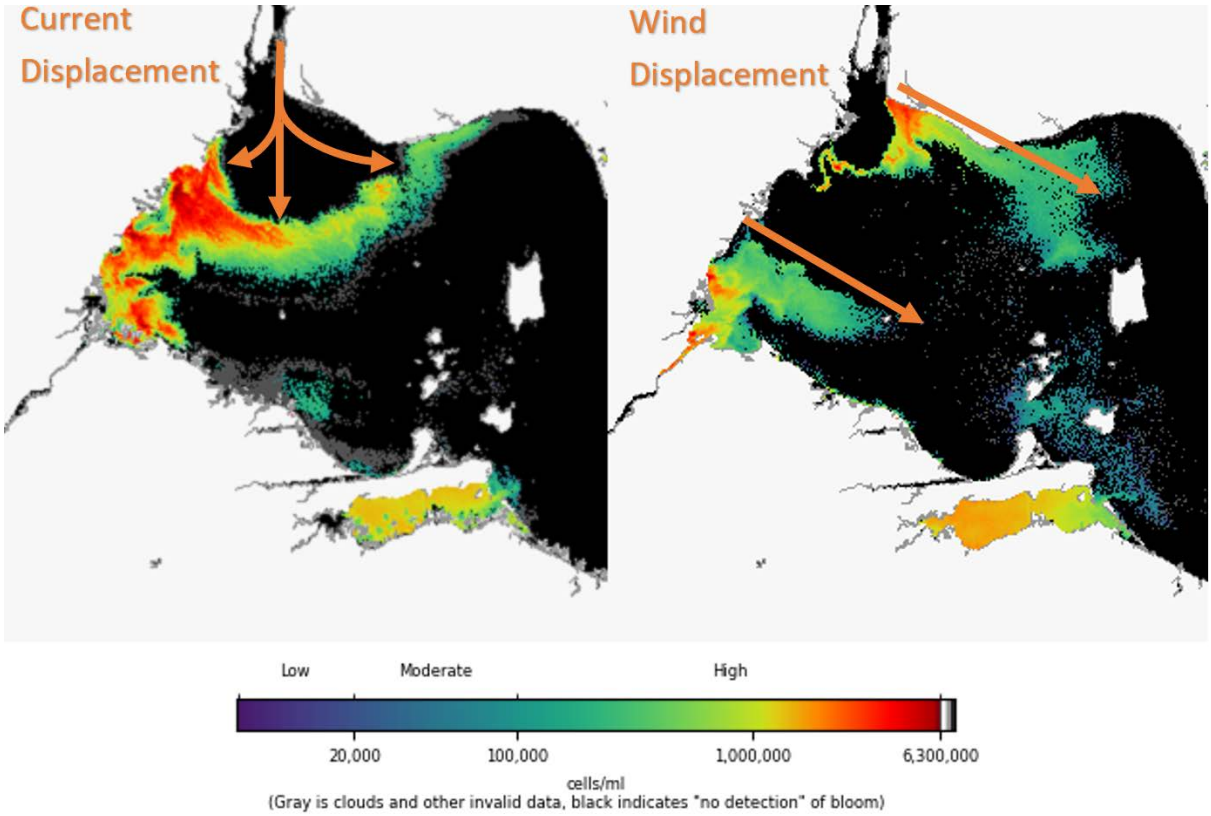

***Figure 6**: Current displacement of Cyanobacteria bloom in 2017, and wind displacement of Cyanobacteria bloom in 2018. Legend signifies Cyanobacteria density in cells/ml.*

We hypothesize that 2016 has produced the most statistically significant results due to having the most stable bloom shapes. Figure 6 reveals bloom shapes for 2017 and 2018 to be stretched and displaced due to wind and current, thus making Cyanobacteria values observed for 2017 and 2018 spatially displaced, therefore less valid. Future models can be improved to account for this by using larger buffer areas to capture displaced blooms, by using data with a higher temporal resolution to capture blooms before they become displaced, by using algorithms to identify a displaced bloom by analyzing it' s shape, and by factoring in a weighting system to account for wind, wind direction, and current of the local area. Additionally, average wind speeds we calculated from our sampling stations for 2016 were lower than that of 2017, and 2018, adding further weight to this hypothesis.

Therefore, we will use the 2016 model to conduct further statistical analysis. When this model is used as the input in RStudio and the ArcGIS OLS regression tool, a number of different statistics are produced as shown in table 5. Figure 7 and 8 both

visualize the distribution of the observed variables and their residuals. Figure 7 shown below uses the regression equation produced by the 2016 model as seen in Table 4, to plot a pink regression plane which symbolizes predicted results. Actual results are plotted by points of varying colour signifying the intensity of Cyanobacteria density as described by the attached legend. The distance between the points and the plane signifies the residuals.

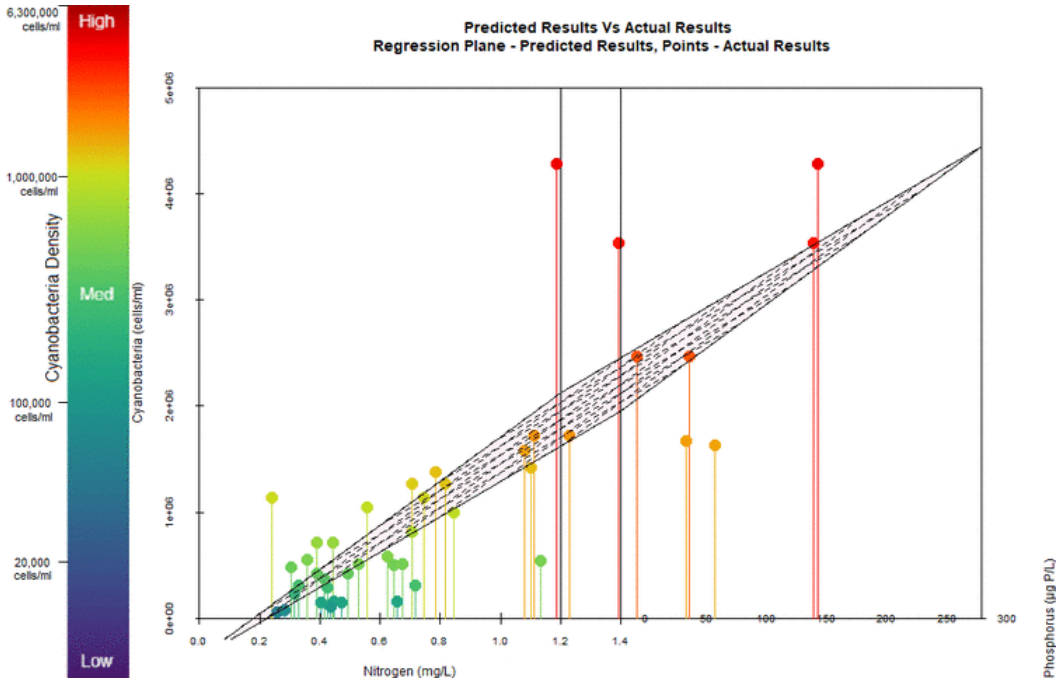

Figure 7: *Graph visualizing actual results versus predicted results of the 2016 regression.*

***Table 5**: OLS Regression Summary and Diagnostics for 2016 with phosphorus and nitrogen*

Summary of OLS Results								
Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
PHOSPHORUS	8307.305887	2151.507799	3.861155	0.000403*	3032.016803	2.739861	0.009140*	2.067325
NITROGEN	1660394.275455	512912.327570	3.237189	0.002428*	811875.459198	2.045134	0.047460*	2.067325

OLS Diagnostics			
Input Features:	2016_Model	Dependent Variable:	CYANOBACTERIA
Number of Observations:	43	Akaike's Information Criterion (AICc) [d]:	1276.195652
Multiple R-Squared [d]:	0.691403	Adjusted R-Squared [d]:	0.675973
Joint F-Statistic [e]:	44.809382	Prob(>F), (2,40) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	77.302484	Prob(>chi-squared), (2) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	4.246405	Prob(>chi-squared), (2) degrees of freedom:	0.119648
Jarque-Bera Statistic [g]:	70.200695	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

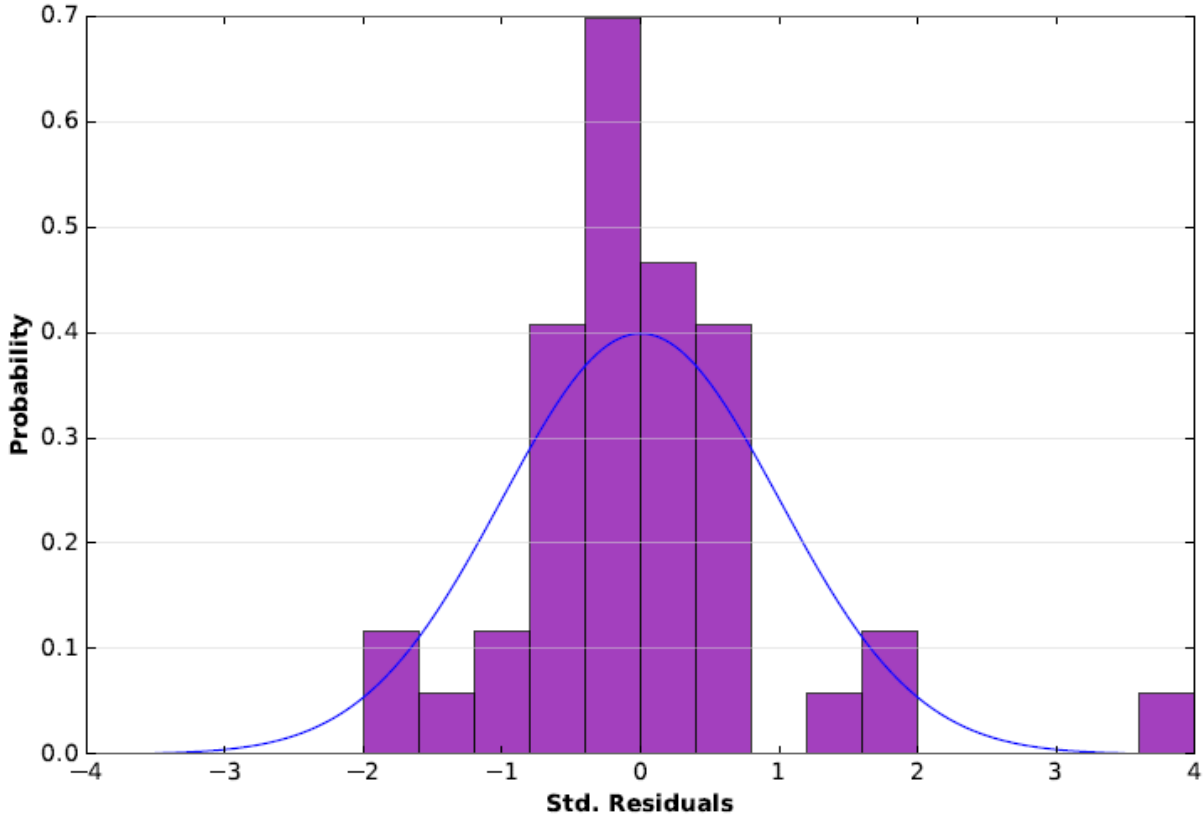

Figure 8: *Histogram of Standardized residuals*

There are several statistics in Table 5 that can be used to analyze the strengths and weaknesses of the model we've produced. One of these statistics is the Koenker statistic. The asterisk next to this value indicates that there is most likely a nonstationary relationship in the model, and performing a Geographically Weighted Regression (GWR) may improve the results of the model. Therefore, we will run the same model, but this time with GWR.

Table 6*: Geographically Weighted Regression Results Summary*

	Weighted Regression
Neighbours	37
Residual Squares	10672371379054.516
Effective Number	8.1908740762399255
Sigma	553711.9618959678
AICc	1269.2377129682613
R^2	0.79170586783759678
R^2 Adjusted	74867643703602826

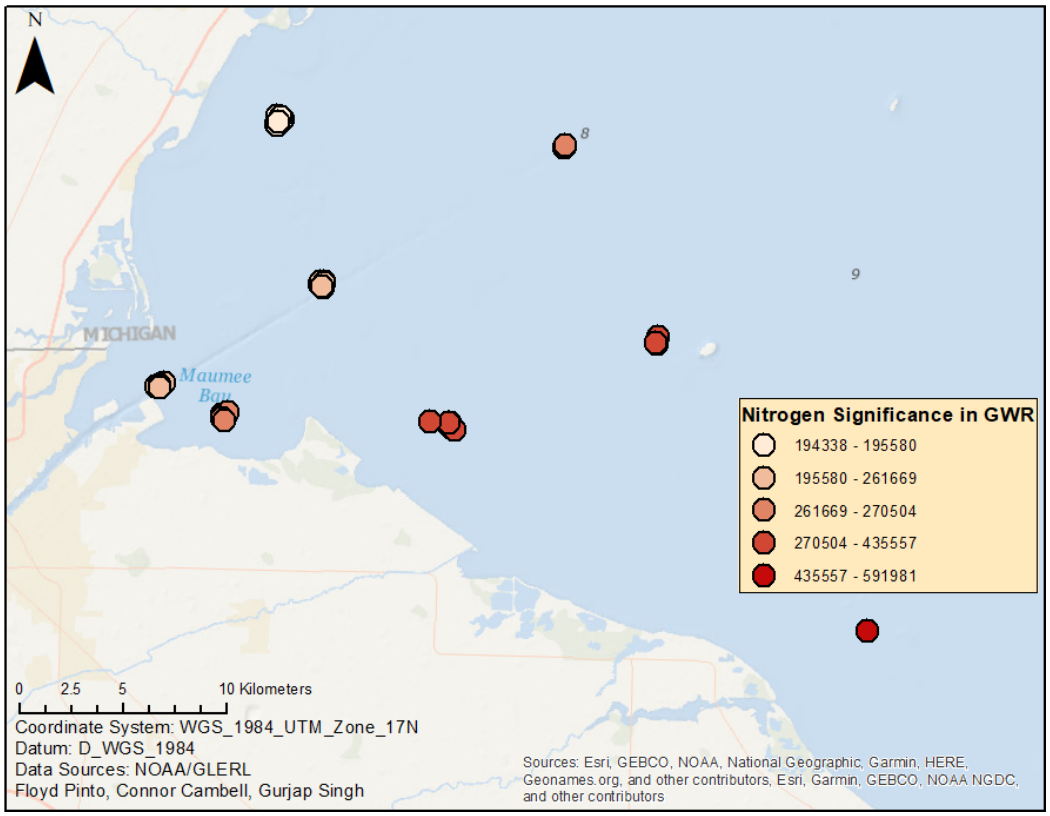

***Figure 9**: Nitrogen Significance in Geographically Weighted Regression*

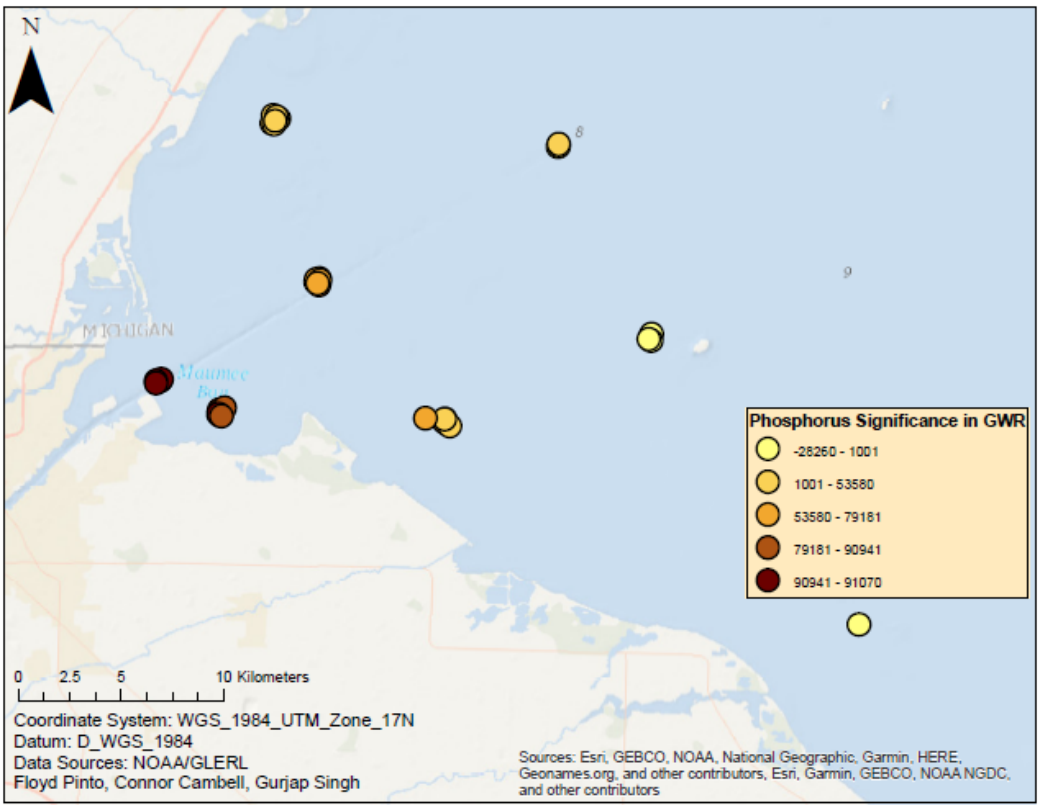

***Figure 10**: Phosphorus Significance in Geographically Weighted Regression*

The results of the GWR in Table 6 show a significant improvement to the results of the model. The Adjusted R-Squared value has increased from about 0.678 to about 0.748, meaning the explanatory variables can now explain about 7% more of the response variable. The AIC value has also decreased from the 1276 value shown in Table 5 to a value of 1269 in Figure 9, proving that the new GWR model does a better job of explaining the response variable than the regular OLS model. Figure 10 shows a high weighting of Phosphorus at Maumee Bay, confirming our hypothesis that this was a problem area due to nutrient runoff. From all this we can conclude that a nonstationary relationship exists between the response variable and the explanatory variables.

Finally, the strengths and weaknesses of the model will be examined by answering the questions that were proposed in the fourth research approach.

1. Are the explanatory variables helping the model?

This relates to the p-values and Koenker statistic from Table 5. The p-values for each variable are below 0.05, so each variable is statistically significant to the model. The Koenker statistic is also statistically significant, but that has already been accounted for with the GWR.

2. Are the relationships what we expected?

As shown in Table 5, there is no negative sign in front of the coefficient for both phosphorus and nitrogen. This indicates they both have a positive relationship with the response variable, which is also observed in Figure 7. As phosphorus and nitrogen levels increase, so does the density of algal blooms. This is the relationship we expected.

3. Is the model biased?

Table 5 shows that our model has a statistically significant Jarque-Bera statistic, meaning the residuals are not normally distributed with a mean of zero.. This means the model is biased, and could be caused by a missing key explanatory variable or the modelling of a nonlinear relationship.

4. Do we have all key explanatory variables?

The results of the spatial autocorrelation show that no clustering occurs in the residuals. This should indicate that there are no missing variables, however the fact remains that the Jarque-Bera statistic is significant. Since the model currently only

examines two explanatory variables for an issue as complicated as algal blooms, it can be assumed that we are most likely missing key explanatory variables.

5. How well are we explaining the response variable?

This relates to both the adjusted R-Squared and the AIC. Through the results shown in Table 6 from the GWR we can confirm that these values indicate a model containing significant relationships. It passes our minimum R-Squared threshold of 0.5, and is roughly at our set forth 5% error margin for average residuals.

This research conducted can have many benefits. These blooms impact several industries such as tourism, fishing, and recreation (Watson et al, 2016). Government agencies monitor blooms to issue safety warnings for activities involving water bodies that have been impacted by blooms (World Health Organization, 2003). Having a predictive model that can show relationships between nutrients, climate change, and Cyanobacterial blooms can make way for potential policy to further constrain the inputs causing Cyanobacteria blooms. From an economic standpoint, this would additionally help by decreasing the impact of Cyanobacterial blooms on revenue loss through tourism, public health risk, and ecosystem disruption (Watson et al., 2016).

7. Conclusion

The purpose of this analysis was to investigate significant relationships between the density of algal blooms in relation to phosphorus, nitrogen and water surface temperature. Knowing what factors contribute significantly to the growth of algal blooms can help develop policies such as best management practices to limit runoff, and safety measures regarding allowable nutrient thresholds to reduce the growth of blooms, and thus the harmful impacts that they may have on the environment. Further literature review revealed that some of the variables that contribute most to algal blooms growth include phosphorus, nitrogen and water temperature. Multiple different regression models were then run against the available data in order to find reliable results. Based on our results, water temperature did not have a statistically significant relationship with the response variable. It was concluded that both phosphorus and nitrogen have a positive linear relationship with Cyanobacteria density. Once the parameters for the final regression model were decided upon, those same parameters were used in a geographically weighted regression to determine whether or not the relationships varied over space. The relationship is also nonstationary, and more heavily weighted at Maumee Bay, making the Maumee River inlet suspect of being the most problematic run-off source.

It is important to remember that the results can only be as accurate as the raw data. Factors such as wind and cloud cover restricted the amount of usable cyanobacteria data. Factors such as the number of sample points and temporal resolution affected the amount of usable nutrient and water temperature data. This study could only use the temporal overlap between the response and explanatory variables, which resulted in a relatively small number of data points compared to what was originally planned. Fewer data points means a smaller data set, and a smaller data set means the data is less representative of the actual scope of information that is present in the real world.

Future improvements to this model would be to factor in bloom displacement by incorporating local wind measurements, current measurements, and analysis of bloom shape. Analyzing the problem using a greater temporal resolution, and tracking the inputs of nitrogen and phosphorus at their source may provide greater cognizance into the relationships at play. The role of water temperature in this model is still underdeveloped and can be further analyzed by taking into account environmental phenomena like aquatic seasonal turnover, and climate change.

References

About the Experimental Lake Erie Harmful Algal Bloom (HAB) Tracker. (n.d.). NOAA Great Lakes Environmental Research Laboratory. https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/habTracker_about.html

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike* (pp. 199-213). Springer, New York, NY.

Altieri, A. H., & Gedan, K. B. (2014). Climate change and dead zones. https://doi.org/10.1111/gcb.12754

Bachmann, R. W., & Jones, J. R. (1974). Phosphorus inputs and algal blooms in lakes. *Iowa State J. Res*, *49*(2), 155-160. https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1248&context=iowastatejournalofresearch#page=70

Banicki, J. J. (2019, July 11) *NOAA, partners predict large summer harmful algal bloom for western Lake Erie*. NOAA. Retrieved April 10, 2020 from https://news.osu.edu/noaa-partners-predict-large-summer-harmful-algal-bloom-for-western-lake-erie/

Bennington-Castro, J. (2015, August 7). Cyanobacteria: Everyday Health. Retrieved January 25, 2020, from https://www.everydayhealth.com/cyanobacteria/guide/

Bosch, N. S., Evans, M. A., Scavia, D., & Allan, J. D. (2014). Interacting effects of climate change and agricultural BMPs on nutrient runoff entering Lake Erie. *Journal of Great Lakes Research*, *40*(3), 581-589. Retrieved January 25, 2020, from https://doi.org/10.1016/j.jglr.2014.04.011

CBC. (2019, August 7). *Lake Erie algal bloom extends into southern Ont. harbour*. Retrieved April 10, 2020, from https://www.cbc.ca/news/canada/windsor/algal-blooms-colchester-toledo-1.5237286

Chen, M., Ye, T. R., Krumholz, L. R., & Jiang, H. L. (2014). Temperature and cyanobacterial bloom biomass influence phosphorus cycling in eutrophic lake sediments. *PloS one*, *9*(3). https://dx.doi.org/10.1371%2Fjournal.pone.0093130

Environmental Protection Agency (2019, December 17). Climate Change and Harmful Algal Blooms. Retrieved January 25, 2020, from https://www.epa.gov/nutrientpollution/climate-change-and-harmful-algal-blooms

Foster, J. M. (2013, November 20). Lake Erie Is Dying Again, And Warmer Waters And Wetter Weather Are To Blame. Retrieved January 25, 2020, from https://thinkprogress.org/lake-erie-is-dying-again-and-warmer-waters-and-wetter-weather-are-to-blame-96956c15f046/

Government of Canada (2018, February) Canada-Ontario Lake Erie Action Stuff: Partnering on Achieving Phosphorus Loading Reductions to Lake Erie from Canadian Sources. Retrieved April 10, 2020 from https://www.canada.ca/en/environment-climate-change/services/great-lakes-protection/action-plan-reduce-phosphorus-lake-erie.html

Haggert, A. (2019, July 24). *Algal blooms to cost Lake Erie tourism economy $110M: study*. CBC. Retrieved January 25, 2020, from https://www.cbc.ca/news/canada/windsor/algal-bloom-economic-cost-lake-erie-1.5221597

Harke, M. J., Steffen, M. M., Gobler, C. J., Otten, T. G., Wilhelm, S. W., Wood, S. A., & Paerl, H. W. (2016). A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, Microcystis spp. *Harmful Algae*, *54*, 4-20. https://doi.org/10.1016/j.hal.2015.12.007

Herrero, A., Muro-Pastor, A. M., & Flores, E. (2001). Nitrogen control in cyanobacteria. *Journal of bacteriology*, 183(2), 411-425.https://doi.org/10.1128/JB.183.2.411-425.2001

Khan, S. (2012, January 2). Regression analysis. Retrieved February 9, 2020, from https://www.slideshare.net/sabakhan16/regression-analysis-10759319

Lane, R. K. (2019, August 14). Lake. Retrieved February 8, 2020, from https://www.britannica.com/science/lake

Mcleod, S. (2019, May 20). P-values and statistical significance. Retrieved February 9, 2020, from https://www.simplypsychology.org/p-value.html

Moore, S. K., Trainer, V. L., Mantua, N. J., Parker, M. S., Laws, E. A., Backer, L. C., & Fleming, L. E. (2008). Impacts of climate variability and future climate change on harmful algal blooms and human health. *Environmental Health*, 7(2), S4. https://doi.org/10.1186/1476-069X-7-S2-S4

Ogashawara, I. (2019). The Use of Sentinel-3 Imagery to Monitor Cyanobacterial Blooms. *Environments*, *6*(6), 60. https://doi.org/10.3390/environments6060060

Parry, M., Parry, M. L., Canziani, O., Palutikof, J., Van der Linden, P., & Hanson, C. (2007). *Climate change 2007-impacts, adaptation and vulnerability: Working group II contribution to the fourth assessment report of the IPCC* (Vol. 4). Cambridge University Press. https://www.researchgate.net/publication/220042209_Climate_Change_2007_Impacts_Adaptation_and_Vulnerability/link/54f59db30cf2eed5d738b3ad/download

Pitcher, G. C., Figueiras, F. G., Hickey, B. M., & Moita, M. T. (2010). The physical oceanography of upwelling systems and the development of harmful algal blooms. *Progress in oceanography*, *85*(1-2), 5-32. https://doi.org/10.1016/j.pocean.2010.02.002

Posch, T., Köster, O., Salcher, M. M., & Pernthaler, J. (2012). Harmful filamentous cyanobacteria favoured by reduced water turnover with lake warming. *Nature Climate Change*, *2*(11), 809-813. https://doi.org/10.1038/nclimate1581

Ray, S. (2019, September 4). *7 Regression Types and Techniques in Data Science*. Analytics Vidhya. Retrieved January 25, 2020, from https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons. https://books.google.ca/books?hl=en&lr=&id=X2Y6OkXl8ysC&oi=fnd&pg=PR5&dq=regression+analysis&ots=sdpTD_oQnv&sig=VaMXx3YcpkJGUQsu-_QnT2-HWGA&redir_esc=y#v=onepage&q=regression%20analysis&f=false

Sivonen, K., & Jones, G. (1999). *Cyanobacterial toxins. Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management* (Vol. 1).

Smith, D. R., King, K. W., & Williams, M. R. (2015). What is causing the harmful algal blooms in Lake Erie?. *Journal of Soil and Water Conservation*, *70*(2), 27A-29A. https://doi.org/10.2489/jswc.70.2.27A

Smith, R. B., Bass, B., Sawyer, D., Depew, D., & Watson, S. B. (2019). Estimating the economic costs of algal blooms in the Canadian Lake Erie Basin. *Harmful algae*, *87*. https://doi.org/10.1016/j.hal.2019.101624

Vander Woude, A., Ruberg, S., Johengen, T., Miller, R., & Stuart, D. (2019). Spatial and temporal scales of variability of cyanobacteria harmful algal blooms from NOAA GLERL airborne hyperspectral imagery. *Journal of Great Lakes Research*, *45*(3), 536-546. https://doi.org/10.1016/j.jglr.2019.02.006

Vijalapuram, Sharad (2019, March 31). *How to read a Regression Table*. freeCodeCamp. Retrieved April 10, 2020, from https://www.freecodecamp.org/news/https-medium-com-sharadvm-how-to-read-a-regression-table-661d391e9bd7-708e75efc560/

Vermaas, W. F. (2001). Photosynthesis and respiration in cyanobacteria. *eLS*. https://doi.org/10.1038/npg.els.0001670.

Watson, S. B., Miller, C., Arhonditsis, G., Boyer, G. L., Carmichael, W., Charlton, M. N., & Matisoff, G. (2016). The re-eutrophication of Lake Erie: Harmful algal blooms and hypoxia. *Harmful algae*, *56*, 44-66. Retrieved January 25, 2020 from https://doi.org/10.1016/j.hal.2016.04.010

World Imagery. [basemap]. January 6, 2020. Scale undetermined; generated by Esri; ArcMap. <https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9> (February 7, 2019)

World Health Organization. (2003). *Guidelines for safe recreational water environments: Coastal and fresh waters* (Vol. 1). World Health Organization.