2017

Forecasting Harmful Algal Blooms for Western Lake Erie Using Data Driven Machine Learning Techniques

Nicholas L. Reinoso
Cleveland State University

FORECASTING HARMFUL ALGAL BLOOMS FOR WESTERN LAKE ERIE USING DATA DRIVEN MACHINE LEARNING TECHNIQUES

NICHOLAS L. REINOSO

Bachelor of Science in Civil Engineering
Cleveland State University
May 2015

submitted in partial fulfillment of requirements for the degree
MASTER OF SCIENCE IN CIVIL ENGINEERING
at the
CLEVELAND STATE UNIVERSITY
May 2017

We hereby approve this thesis for

Nicholas L. Reinoso

Candidate for the Master of Science in Civil Engineering degree for the

Department of Civil and Environmental Engineering

and the CLEVELAND STATE UNIVERSITY'S

College of Graduate Studies by

Dr. Ungtae Kim

Department & Date

Dr. Mehdi Jalalpour

Department & Date

Dr. Walter Kocher

Department & Date

5/3/2017

Student's Date of Defense

DEDICATION

I dedicate this thesis to my parents for their endless love, support, and encouragement.

ACKNOWLEDGEMENT

I would like to thank Dr. Ungtae Kim for his support, encouragement, and his invaluable advice throughout this study. I would also like to express my sincere thanks to committee members, Drs. Walter Kocher and Mehdi Jalalpour for their valuable comments. I also thank Cleveland State University, Washkewicz College of Engineering, and the College of Graduate Studies for providing me with the opportunity to pursue my thesis. This thesis was supported by Dr. Ungtae Kim's Startup Fund from Cleveland State University.

A special thanks to my family and friends especially Seema and Amin for working alongside me and supporting me these last two years. I am also thankful for my brothers, Mike and Joe, whose support has never wavered. I am forever grateful.

FORECASTING HARMFUL ALGAL BLOOMS FOR WESTERN LAKE ERIE USING DATA DRIVEN MACHINE LEARNING TECHNIQUES

NICHOLAS REINOSO

ABSTRACT

Harmful algal blooms (HAB) have been documented for more than a century occurring all over the world. The western Lake Erie has suffered from Cyanobacteria blooms for many decades. There are currently two widely available HAB forecasting models for Lake Erie. The first forecasting model gives yearly peak bloom forecast while the second provides weekly short-term forecasting and offers size as well as location. This study focuses on bridging the gap of these two models and improve HAB forecast accuracy in western Lake Erie by letting historical observations tell the behavior of HABs. This study tests two machine learning techniques, artificial neural network (ANN) and classification and regression tree (CART), to forecast monthly HAB indicators in western Lake Erie for July to October. ANN and CART models were created with two methods of selecting input variables and two training periods: 2002 to 2011 and 2002 to 2013. First a nutrient loading period method which considers all nutrient contributing variables averaged from March to June and second a Spearman rank correlation to choose separate input sets for each month considering 224 different combinations of averaging and lag periods. The ANN models showed a correlation coefficient increase from 0.70 to 0.77 for the loading method and 0.79 to 0.83 for the Spearman method when extending the training period. The CART models followed a similar trend increasing overall precision from 85.5% to 92.9% for the loading method and 82.1% to 91% for the

Spearman method. Both selection methods had similar variable importance with river discharge and phosphorus mass showing high importance across all methods. The major limitation for ANN is the time required for each forecast to be complete while the CART forecasts earlier is only able to produce a class forecast. In future work, the ANN model accuracy can be improved and use new sets of variables to allow earlier HAB forecasts. The final form of ANN and CART models will be coded in a user interface system to forecast HABs. The monthly forecasting system developed allows watershed planners and decision-makers to timely manage HABs in western Lake Erie.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Page

ACRONYMS

Air	Air Temperature
ANN	Artificial Neural Network
CART	Classification and Regression Tree
Chl-a	Chlorophyll-a
CI	Cyanobacterial Index
EPA	Environmental Protection Agency
GLERL	Great Lakes Environmental Research Laboratory
GLM	Great Lakes Monitoring
GLNPO	Great Lakes National Program Office
GP	Genetic Programming
HAB	Harmful Algal Bloom
NCWQR	National Center for Water Quality Research
NOAA	National Oceanic and Atmospheric Administration
PM	Phosphorus Mass
ppb	Parts Per Billion
Q	River Discharge
SRP	Soluble Reactive Phosphorus
TKN	Total Kjeldahl Nitrogen
TP	Phosphorus Concentration
USGS	United States Geological Survey
Water	Water Temperature
Wind	Wind Speed
WLEEM	Western Lake Erie Ecosystem Model

CHAPTER I

INTRODUCTION

1.1 Problem Statement

Harmful Algal Blooms (HAB) are quickly becoming a major problem all around the world. A HAB is a bloom of algae that has the potential to harm humans or the ecosystem (Ho, 2015). The HAB problem is well documented impacting recreation, water treatment, individual health, and property values. The species of HABs in western Lake Erie is Microcystis where bloom growth is promoted by warm temperatures over twenty degrees Celsius. The months that consistently have temperatures over the temperature threshold are July, August, and September. The months that often have blooms are the three months over the temperature threshold with a carry over into October. HABs are being forecasted by different techniques around the world.

Machine learning techniques have been increasingly used to forecast HABs. Dissimilar to traditional methods, machine learning is based on algorithms that are able to iteratively learn from data finding hidden insights without depending on rule-based programming. Supervised learning algorithms are often used when historical data is able to predict future events.

This study uses two supervised machine learning techniques with two different methods of choosing input data for two training periods for a total of eight models briefly described as follows. The first technique used is classification and regression tree (CART) forecasting for the severity of the bloom with two considered ranges, classes 1 to 3 and 1 to 5. The second technique used is artificial neural networks (ANN) forecasting the biomass of the bloom. Two methods of selecting input variables are used for both techniques and both training periods. The first method is an accepted nutrient loading period determined through the literature review using nutrient contributing variables. The second method used for selecting the input sets is the Spearman rank correlation which also considers variables that affect the growth of HABs such as temperature and wind speed as well as the nutrient contributing variables. Both techniques use two training periods from 2002 to 2011 and from 2002 to 2013.

Forecasting HABs in Lake Erie will allow commercial as well as recreational users of the lake to make timely decisions concerning western Lake Erie. There are two available HAB forecasting models for western Lake Erie from the National Oceanic and Atmospheric Administration (NOAA). One of the forecasts is an assembly of multiple models to forecast the peak bloom for the year. The second forecast is focused on weekly short-term forecasting and provides size as well as location. The focus of this study is to bridge the gap between the two available forecasts.

1.2 Study Area

This study focuses on forecasting HABs in the western basin of Lake Erie. The western basin of Lake Erie has had a problem with HABs for decades whereas the central and eastern basins have not experienced large HABs. There are two major factors that cause this to occur: water depth and nutrient loading. The average water depths for the three basins are 7.4, 18.3, and 24 meters for the western, central, and eastern basins, respectively. The shallow waters in the western basin cause an increase in water temperature promoting the growth of HABs. Figure 1 below shows the average monthly water temperature for the western Lake Erie basin.

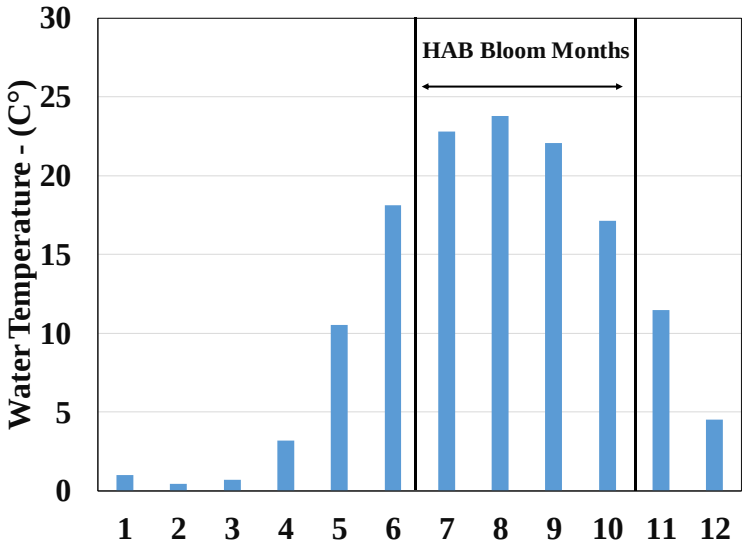


Figure 1. Monthly Average Water Temperature in the Western Lake Erie Basin from 2000 to 2015

The second major factor that promotes bloom growth in the western basin is nutrient loading. The two major tributaries into Lake Erie are the Detroit River and the Maumee River. The two main nutrients for Microcystis to bloom are phosphorus and nitrogen with phosphorus being the limiting factor. The amount of flow from the Maumee River is

1/35th of the Detroit River however the concentration of nutrients results in the same amount of nutrients entering the lake from the Maumee River (Stumpf, 2016). The remaining tributaries are insignificant producing less than ten percent of the nutrient loads of the Maumee River (Stumpf, 2016). In this study through the literature review, it was determined that the nutrient loads from the Maumee River are the main source of nutrients for modeling HABs. Figure 2 below shows the study area.

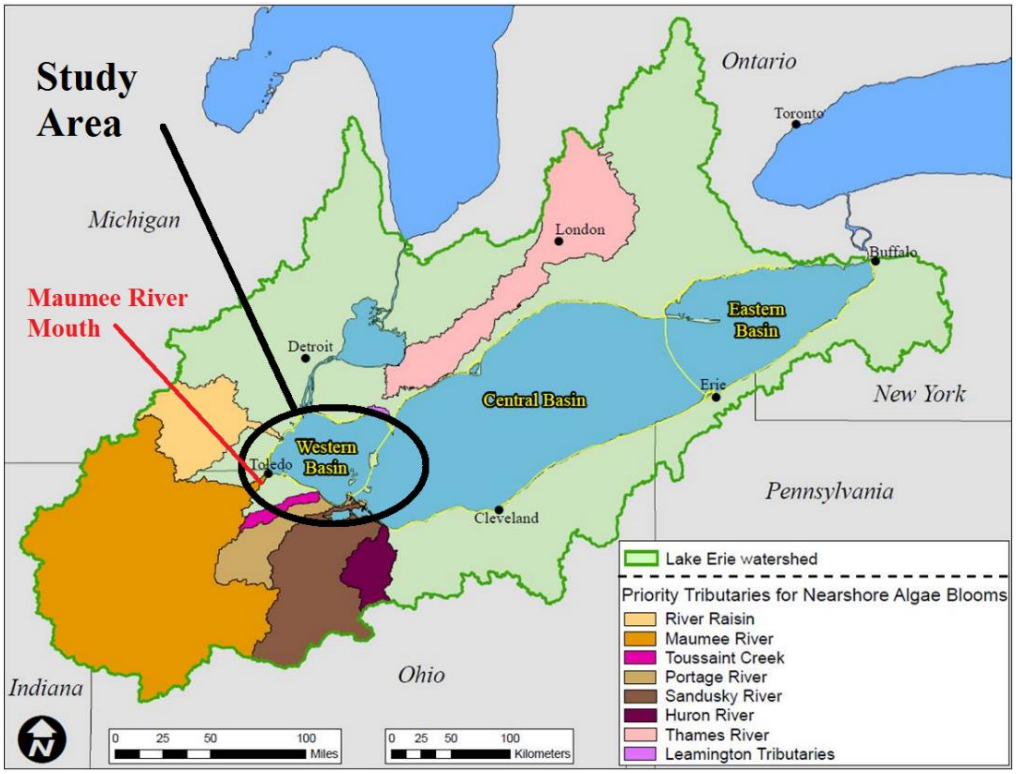


Figure 2. Map of the Western Lake Erie Basin with Priority Tributaries for HABs (Environmental Protection Agency [EPA], 2017)

Lake Erie's problem with HABs has been going on for decades and has been well documented since the 1950s. By the mid-1960s Lake Erie was declared "dead" and HABs were reported seasonally in the western basin of Lake Erie. The HABs were driven by the large amounts of phosphorus and nitrogen entering the lake from sources such as

farm runoff, sewage, and industrial pollution. In 1972, an effort to fix the lake began. The Clean Water Act was passed in 1972 to increase the regulations on industrial dumping. Also, in 1972 the United States and Canada signed the Great Lakes Water Quality Agreement in an effort to reduce the amount of pollutants entering the Great Lakes. In the agreement, the two countries agreed to reduce the amount of phosphorus load entering Lake Erie to 14,600 metric tons from 29,000 metric tons which was later agreed to further reduce the loads to 11,000 per year in a 1978 Agreement (EPA, 2017). By 1977 the Detroit Wastewater Treatment Plant reduced the amount of phosphorus put into the Detroit River by over 90% (Bingham, 2015).

The regulations from 1972 started having an effect and had revived Lake Erie by the 1980s. During the 1980s, no large blooms were reported however in the late 1980s zebra and quagga mussels started to arrive in Lake Erie. In the mid-1990s, large HABs started forming once again. A recent study was performed to look at the reason for the resurgence of HABs. The following list shows some of the twenty-five possible factors examined in the study (Smith, 2015):

- Climate Change – The total amount of rainfall in the HAB loading period has increased roughly 25% in the last decade
- Commodity Prices – Prices for farm goods has been increasing recently which gives producers more of a reason to use more phosphorus to reduce the chance of crop loss
- Fertilizer Source – In the 1990s there was a switch to a fertilizer with increased soluble phosphorus

- Fertilizer Timing – Fertilizer is often applied before crops are grown and with the reduced intake of fertilizer initially there is an increase chance for phosphorus runoff
- Larger Farms – In the last thirty years the number of farms harvesting corn grain has nearly halved resulting in larger farms which now need to apply fertilizer earlier than in previous decades
- No-Till – The no-till growing method was adopted in the 1990s and resulted in an increase phosphorus load through subsurface drainage

During the 2000s, HABs had returned to being a yearly problem for the western Lake Erie basin. Another possible factor for the return of HABs is the zebra and quagga mussels. The mussels filter small particles out of the water such as algae, microscopic bugs, or zooplankton that eat algae (Ruetter, 2014). They then excrete dissolved phosphorus, a main source of food for HABs. If the mussels suck in a harmful form of algae, they stop filtering and spit it out then start filtering again (Ruetter, 2014). The problem of HABs has been becoming an increasingly larger problem for Western Lake Erie. Satellite images have been analyzed by NOAA since 2002 in order to determine the magnitude of the HAB biomass in ten day intervals known as the Cyanobacterial Index (CI). The 2011 bloom peak was 274% larger over the previous peak bloom of the previous nine years. as shown below in Figure 3.

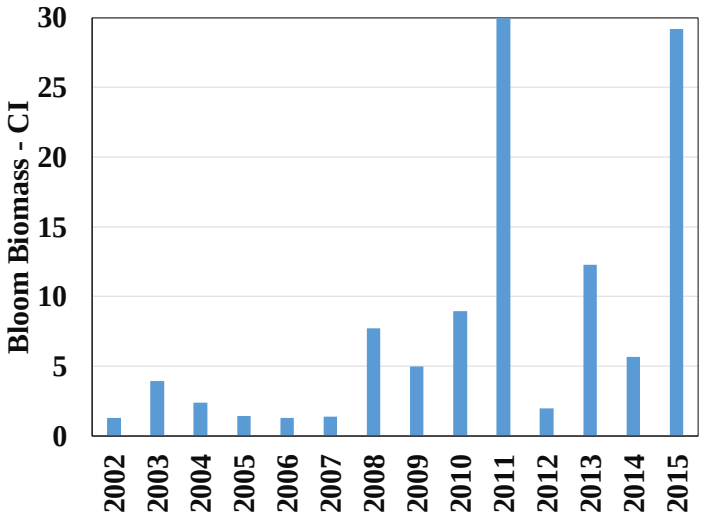


Figure 3. Annual HAB Peak Biomass from 2002 to 2015

Western Lake Erie has had one serious catastrophe in recent times. On August 2, 2014, the City of Toledo's water treatment plant was shut down until August 4th. The bloom was not large in terms of coverage throughout the lake however the bloom was very thick and happened to be concentrated where the water treatment plant's intake pipes are located. When the water in Lake Erie was tested the Microcystin toxin levels were between ten to twenty parts per billion (ppb) (Kozacek, August 2014). The World Health Organization has set the following guidelines for Microcystin in Ohio: children under six and sensitive populations do not drink when the toxin levels reach 0.3 ppb, ages six and older when there is a concentration level of 1.6 ppb, and when the toxin levels reach twenty ppb the water should not be used (EPA, 2017). The drinking water crisis left more than four hundred thousand people and three counties in Ohio and one in Michigan without drinking water (Kozacek, August 2014). The governor of Ohio, John Kasich in 2014, announced a state of emergency to organize resources for the affected and the following emergency measures also became apparent (Kozacek, August 2014):

- Stores sold out of bottled water, sending residents into neighboring cities and states to find supplies
- Local restaurants, universities and public libraries closed
- Several nearby municipalities that were not affected by the toxin offered water to Toledo residents free of charge
- The National Guard was charged with delivering 300 cases of bottled water from Akron, Ohio, as well as Meals Ready to Eat for distribution to homeless shelters and other vulnerable populations who were unable to cook with their water
- Humanitarian organizations like the American Red Cross responded by manning water distribution centers and provided water delivery assistance to homebound residents

1.3 Research Objectives

The main objective of this thesis is to create two models to improve operability and accuracy for forecasting monthly western Lake Erie HAB indicators for July, August, September, and October. Two machine learning techniques were used to create monthly forecasts, ANN and CART. Before this main objective could be accomplished the following three sub-objectives had to be completed:

- Performing a literature review
- Collecting relevant data
- Analyzing and systematically selecting sets of input variables

1.4 Scope and Organization of Thesis

The first step of this study was to perform a literature review to understand the history of HABs in Lake Erie and to examine the current forecasting methods used in Lake Erie and around the world. The second step of the study was to collect relevant data to forecasting HABs from various sources. The third step was to analyze the collected data to determine the importance of each variable and then systematically select sets of input variables. The final step of this study was to use the selected variables to create ANN and CART monthly forecasting models.

This thesis is organized in five chapters. Chapter II goes through the effects of HABs, a review of machine learning models for forecasting, and other work being done on the HAB problem in Lake Erie. Chapter III goes through the collection of meaningful variables for the forecast of HABs, the systematic selection of sets of variables, the forecasting models used, and input variable selection methods. Chapter IV goes through the results of the CART and ANN models and provides a discussion on the results. Chapter V summarizes the study and the key conclusions.

CHAPTER II

LITERATURE REVIEW

This chapter first goes through the effects of HABs. This is followed by the advantages and disadvantages of three machine learning techniques and concludes with an examination of various HAB forecasting models currently available for Lake Erie.

2.1 Harmful Algal Bloom Effects

HABs have the possibility of causing many different types of health problems for humans and animals as well as having major effects on the economy. The most common species of harmful algae in Ohio lakes is Cyanobacteria also known as blue-green algae. The Ohio Department of Health listed the health problems that go along with each type of exposure listed below (Ohio Department of Health, 2016):

a. Drinking or swallowing water contaminated with Cyanobacteria
    - Severe diarrhea and vomiting
    - Difficulty breath
    - Neurotoxicity (weakness, tingly fingers, numbness, dizziness)
    - Death

b. Skin Contact often from recreation activities in HAB waters

- Rashes
- Hives
- Skin blisters

c. Inhaling water droplets of mists of Cyanobacterial contaminated water

- Runny eyes and nose
- Sore throat
- Asthma-like symptoms

HABs can have serious effects on local economies such as property values in western Lake Erie. A study performed to look at the economic effects of HABs determined there is 3.458 billion dollars in residential housing stock near the western basin of Lake Erie (Bingham, 2015). Recreational activities such as boating, skiing, fishing, or swimming are all effected when HABs occur. Water treatment plants have taken more precautions and now use more treatment methods in an attempt to not repeat what happened in Toledo in 2014. Tourism is also an important economic factor; millions of trips are taken to counties near western Lake Erie with a range of sixty-six million to three hundred and five million dollars at risk (Bingham, 2015). Table 1 shows the result of the study on economic loses from the 2011 and 2014 HABs.

Table 1. Breakdown of HAB Impacts on the Ohio Economic Losses (Bingham, 2015)

Economic Factors	**HAB Event Year**	
	2011	**2014**
Property Value	$16,000,000	$18,000,000
Tourism	$20,000,000	$20,000,000
Recreation	$31,000,000	$23,000,000
Water Treatment	$4,000,000	$4,000,000
Overall	$71,000,000	$65,000,000

The Toledo water-treatment plant is in the process of being upgraded. In 2012 the city began working on upgrading the water-treatment plant originally estimated at $312 million over 20 years (Messina, 2016). Residents of Toledo and the surrounding suburbs have seen an increase in their water bills. By 2018 the residents will have an additional hundred and twenty-five dollars on their bill annually when compared to 2013 (Messina, 2017). The Ohio EPA recently mandated that the work be completed in ten years instead of the original plan of twenty years and included a list of mandated additional upgrades that must be completed for an additional $188 million for a total of $500 million including an $80 million upgrade to address HABs (Messina, 2016).

2.2 Harmful Algal Bloom Modeling

2.2.1 Machine Learning Forecasting Techniques

Since the 1990s, machine learning has been used to solve many complicated problems in various fields. Machine learning is an area of computer science and a sub-area of artificial intelligence concentrating on theoretical foundations (Muttil, 2006). Machine learning, in general, contains algorithms that estimate dependency between a systems inputs and outputs while improving its performance automatically through a training period. These different methods are then able to predict outputs from given inputs. These techniques are ideally suited to model the HAB dynamics since such models can be set up rapidly and are known to be effective in handling dynamic, non-linear and noisy data, especially when underlying physical relationships are not fully understood, or when the required input data needed to drive the process-based models are not available (Muttil, 2006). Three artificial intelligence algorithms are examined in this

literature review: ANN, CART, and genetic programming (GP). There are three major tasks shown below that artificial intelligence techniques are used for with the strengths of each algorithm examined in a study from (Kim, 2009) (Figure 4):

- Knowledge engineering which is the process of acquiring knowledge and refining it to gain additional knowledge
- Problem solving such as scheduling and optimization
- Classification and prediction

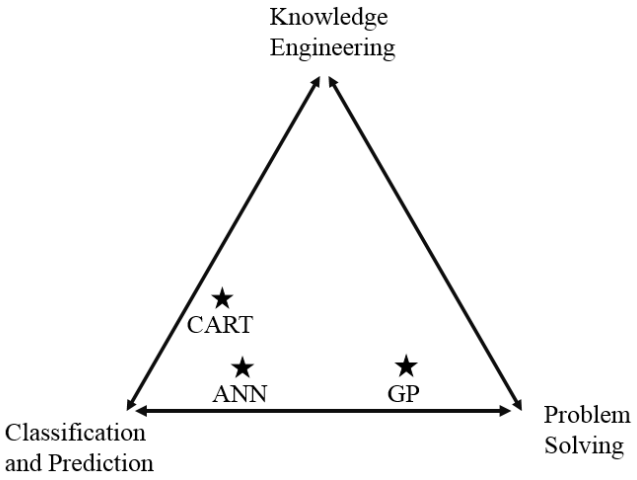


Figure 4. Strengths of Each of the Three Algorithms for Three Major Tasks

Recently machine learning techniques such as ANN and CART have progressively been used to create models for forecasting HABs. ANN models have been used in ecological and environmental science since the 1990s. ANNs can be applied to the following types of problems: pattern classification, clustering and categorization, function approximation, prediction and forecasting, optimization, associative memory, and process control (Kim, 2009).

ANNs are data driven and the size of the sample size can determine the accuracy of the model. ANNs are able to learn patterns and concepts directly from inputs and adjust weights on each neuron even with unknown data relationships and non-linear data (Kim, 2009). ANNs have many advantages over other modeling techniques. One of the largest advantages of ANN's approach as opposed to other models is its ability to deal with uncertain information and incomplete or inconsistent data which makes it good for the forecasting of HABs (Velo-Suarez, 2007). An additional advantage ANN has is that it makes no early assumptions of the network. ANN is also able to handle missing data as well as noisy and scattered data. One of the disadvantages of ANNs is that they are non-transparent black-box models and do not give any exact equations. Another disadvantage is the number of hidden layers and neurons must also be selected by the modeler and can have a considerable influence on performance on the outcome (Pal, 2003). In conclusion, ANNs are good when the results of the model are more imperative and little importance is placed on how the output is determined.

Machine leaning algorithms such as CARTs have been around for decades. The CART algorithm builds classification trees for categorical variables and regression trees for continuous dependent variables (De'ath, 2000). The algorithm repeatedly splits data into two mutually exclusive groups by following simple rules until more branching would not add any accuracy to the output. CART has many advantages compared to other models. Unlike ANN models, it is not black box as it creates easily understandable rules. CART generally takes less overall time to train and create models and is able to easily rank the importance of input variables when compared to other machine learning

techniques such as ANN and GP, (Kim, 2009). According to (De'ath, 2000) CART models also have the following advantages:

- The flexibility to handle a broad range of response types, including numeric categorical, ratings, and survival data
- Invariance to monotonic transformations of the explanatory variables
- The ability to handle missing values in both response and explanatory variables

CART is non-parametric and is able to discover complex interactions between inputs which could be challenging to determine using traditional multivariate methods (Lewis, 2000). Another major advantage of CART is it is able to scale to large problems and is able to better model small data sets compared to ANN (Razi, 2005). CART also has disadvantages attached to it. CART is unable to do various functions such as expressing linear relationships easily, produce unique solutions, or produce continuous outputs due to its binary nature (Kim, 2009).

For the ANN and CART models to find their optimal results, a variable selection analysis is often performed. When many lag and averaging periods are considered for each input variable, many of the variables can possibly have no significant effect on the output. The presence of many irrelevant variables in the ANN model can result in the model behaving poorly. A study was performed on the selection of input variables for ANNs for water resources applications, the five methods overviewed in the study are listed below (Bowden, 2005):

- Selection of variables and the corresponding lag and averaging periods based on knowledge of the system

- Selection of variables based on the correlation between inputs and outputs with all considered lags and averaging periods
- Selection of variables through a heuristic method of trial and error testing different sets of input variables in an ANN model or through the use of stepwise selection
- Selection of variables by examining an ANN model with all considered input variables and determining the relative importance of each variable
- Selection of variables through a combination of the four other methods

Input variable selection is an important part of machine learning model development due to the negative impact that poor selection can have on the performance during training and deployment post-development (May, 2011). Selecting the optimal set of input variables before creating machine learning models reduces the computational strain and overall effort related to training and selecting the final model (May, 2011). (Zeng, 2010) recently researched an ANN pre-warning HAB model for a lake in Beijing with three classes, green, yellow, and red, for chlorophyll-a (Chl-a) values greater than 0.10 mg/L between 0.06 mg/L and 0.10 mg/L, and less than 0.06 mg/L. The three classes were determined through the hydrographic history and historical experiences. The first step of the study was to collect water quality, meteorological, and hydrological historical data from the lake. The second step of the study was analyzing the correlations of nitrogen and phosphorus against Chl-a to determine which variable to represent nutrients. (Lee, 2003) researched ANN modeling for HABs for the coastal waters of Hong Kong. Through field and modeling studies, ten input variables were chosen for forecasting the HABs. The

study constructed numerous models to determine the optimal set of input variables considering seven different time lags.

Since the 1960s, GP has been used to solve problems and has recently been applied to forecasting HABs. (Sivapragasam, 2010) recently described how genetic programing is similar to genetic algorithms as it is an evolutionary algorithm based on Darwinian theories of natural selection and survival of the fittest. The algorithm begins with an initial set of random equations which can include arithmetic operators (plus, minus, multiply, and divide), mathematical functions (sin, cos, exp, and log) and logical functions (or, and) (Sivapragasam, 2010). This set of possible solutions is then tested using the GP algorithm, after which the equations that best fit the training data are selected. The (Sivapragasam, 2010) study referenced how the transparent nature of GP solutions may allow inferences about underlying processes to be made, highlighted issues with scaling data for machine learning and noted the difficulty involved with producing understandable models.

The use of GP in forecasting HABs is not without its advantages and disadvantages. One disadvantage of using the GP algorithm is that the user must decide a number parameters before applying the algorithm to model the data, such as number of equations and number of calculation generations. The main advantage of GP is its ability to produce models that build a definitive formula or equation. Study results from (Razi, 2005) show that "ANNs and CART models provide better prediction compared to regression models when the predictor variables are binary or categorical and the dependent variable continuous."

2.2.2 Current Harmful Algal Bloom Forecasting Techniques

Two different widely available forecasts for HABs in Lake Erie are accessible to the public through NOAA's web site. One of these forecasts predicts the peak bloom for the entire year while the other tracks and gives a 5-day forecast on the growth and movement of the bloom. The two models are updated weekly and daily, respectively. The forecast that predicts the peak bloom is a combination of four models: NOAA-Q, NOAA-TBP, University of Michigan / North Carolina State University / NOAA Great Lakes Environmental Research Laboratory (GLERL) - Bayesian, and Limnotech – Western Lake Erie Ecosystem Model (WLEEM). The five-day forecast/tracker is developed by NOAA, GLERL, and Cooperative Institute for Limnology and Ecosystems Research.

A study performed by (Stumpf, 2012) that observed HABs through NOAA satellite imagery determined a CI based on the biomass of the blooms. In the study, it was determined that the blooms often peak in August or September and are correlated to discharge as well as the phosphorus load from the Maumee River. The NOAA-Q and NOAA-TBP are empirical statistical-heuristic models that use flow discharge and total bioavailable phosphorus from the March to June nutrient loading season. The WLEEM model is a process-based fine-scale 3D linked hydrodynamic-sediment transport-advanced eutrophication model. The Bayesian model is an empirical Bayesian model relating spring phosphorus loading to multiple estimates of HAB size. The Bayesian model forecast relates bloom size to spring phosphorus loads, as well as considering an increase of susceptibility to HABs. The model is calibrated to the CI algorithm data developed by (Stumpf, 2012).

The yearly peak bloom forecast is first announced at a webinar presented by the Ohio Sea Grant and is updated weekly on a NOAA bulletin. The forecast is presented in a bloom severity index ranging from 0 to 10 with a five-class breakdown: Class 1 (0 to 2), Class 2 (2 to 4), Class 3 (4 to 7), Class 4 (7 to 9), and Class 5 (>9). The HAB Tracker is a tool that combines remote sensing, monitoring, and modeling to produce daily 5-day forecasts of bloom transport and concentration. The HAB Tracker looks at daily satellite images and real-time data and estimates the current size and intensity of the HAB. The forecasting part of the tracker uses forecasted meteorological data and hydrodynamic modeling to forecast where the bloom will travel as well as the concentration of the bloom.

There are advantages and disadvantages of the two different forecasts available. An advantage for the HAB tracker includes predictions which aid public health officials and water intake managers in a making timely public health decisions. One of the main advantages is that the tracker gives live information on the exact intensity, size, and location where an HAB is occurring. This information gives recreational users the condition of the water for lake activates and city managers information to create public health decisions. One of the disadvantages is that with only a five-day forecast when a large bloom is imminent it does not give city officials enough time to prepare. The yearly peak bloom forecast also has advantages and disadvantages. The main advantage is that if a large boom is forecasted then everyone on Lake Erie is able to prepare. For example, when a large bloom is predicted for the year cities such as Toledo may stock pile water bottles in case the bloom causes the water plants to be shut down. An annual peak forecast disadvantage is that it is unknown when the bloom will actually happen meaning

the decisions are based only on HAB peaks which typically occur in August or September. Another disadvantage is that there is no information forecasted on where the bloom will occur and which cities and areas it will affect.

Creating data driven monthly forecasting models will be useful when used in combination with the other already created models. Being able to forecast each month individually has advantages for decision makers. With monthly forecasts, we will be able to predict for each month whether a bloom will occur and, if so, the class of bloom that will happen ranging from Class 1 to 3 or 1 to 5. Another advantage is that more variables will be considered over other models such as wind speed and temperature. One of the disadvantages to the model is that it does not forecast exactly where the bloom will occur; however, this problem is partially fixed when the model is used in combination with NOAA's HAB tracker when the HAB starts to arise.

CHAPTER III

METHODS

This chapter first details the variables that were considered for forecasting HABs as well as the gathering of the data. This is followed by an analysis of the variables and an overview of the forecasting models used in this study. Lastly, this chapter goes through the selection criteria of the input sets after the initial set of variables was narrowed down.

3.1 Data Gathering

One main objective of this study was to gather numerous independent and dependent variables to be considered. Initially, there was a total of nineteen independent and six dependent variables considered shown in Tables 2 and 3 below. The first six variables were obtained from the Heidelberg Tributary Loading Program operated by Heidelberg University's National Center for Water Quality Research (NCWQR) (Heidelberg, 2017). Water samples were taken on the Maumee River at Waterville, OH at the United States Geological Survey (USGS) station (04193500), one to three samples are analyzed a day depending on times of high flow or turbidity. The ten variables were taken from Great Lakes Monitoring (GLM) from the Illinois-Indiana Sea Grant. The CI data was gathered

from NOAA and Dr. Stumpf (Stumpf, 2016). Chl-a data was taken from GLM and EPA's Great Lakes National Program Office (GLNPO). Two satellites were used: the Medium Resolution Imaging Spectrometer for 2002-2011 and the Moderate Resolution Imaging Spectroradiometer for 2012-2015 (Stumpf, 2016). Ten-day composite images of the maximum CI at each map pixel were determined by using the satellite images to determine the total biomass for the ten-day periods from July 11th to October 31st (Stumpf, 2016). After collecting these ten-day CI values, they were converted to the max value of the month as well as the average of the CI values in each month.

Table 2. List of All Considered Input Variables.

Variable		Unit	Method	Source
Q	River Discharge	m^3/s (cms)	Average	NCWQR
TP	Phosphorus Concentration	mg/L	Average	NCWQR
PM	Phosphorus Mass	Ton	Average	NCWQR
SRP	Soluble Reactive Phosphorus	mg/L	Average	NCWQR
TKN	Total Kjeldahl Nitrogen	mg/L	Average	NCWQR
S	Sulfate	mg/L	Average	NCWQR
Q	River Discharge	Cms	Average	GLM
TP	Phosphorus Concentration	mg/L	Average	GLM
SRP	Soluble Reactive Phosphorus	mg/L	Average	GLM
TKN	Nitrogen Concentration	mg/L	Average	GLM
T	Turbidity	NTU	Average	GLM
A	Alkalinity	mg/L	Average	GLM
N-N	Nitrite-Nitrate	mg/L	Average	GLM
TP	Total Phosphorus	ug/L	Average	GLM
DO	Dissolved Oxygen	mg/L	Average	GLM
CI	Cyanobacterial Index	10^{20}	Average, Max	NOAA
Water	Water Temperature	C°	Average	USGS
Air	Air Temperature	C°	Average	USGS
Wind	Wind Speed	Knots	Average	USGS

Table 3. List of All Considered Dependent Variables.

Variable		Unit	Method	Source
Chl-a	Chlorophyll a	ug/L	Average	GLM
Chl-a	Chlorophyll-a	ug/L	Average of Max	EPA GLNPO
Chl-a	Chlorophyll-a	ug/L	Max of Max	EPA GLNPO
CI	Cyanobacterial index	10^{20}	Average	NOAA
CI	Cyanobacterial index	10^{20}	Max	NOAA
Mc	Microcystin	ug/L	Average	GLERL

The importance and impact on HABs for the variables was researched in the literature review phase to determine the key variables. Phosphorus as well as nitrogen are the two main sources of nutrients for the HABs in Lake Erie. One of the most important variables that promotes bloom growth is phosphorus which is often the nutrient that there is less of in freshwater whereas nitrogen is the limiting nutrient factor in saltwater. Phosphorus is a crucial nutrient for life forms and is required for metabolic reactions in plants to grow (Lawson, 2011). River discharge from the Maumee River is an important variable in determining the amount of nutrients entering Lake Erie. Low turbidity normally due to slow moving water allows more light to penetrate through the water column creating optimal conditions for HABs to grow (Indiana University, 2017). Temperature is also an important factor in the growth of Cyanobacteria, When the water temperature is over twenty-five degrees Celsius, it is an optimal time for growth (Indiana University, 2017). Wind speed and direction also play a part in HABs. The speed affects the size and intensity of the HABs whereas the direction can determine where the bloom travels. Low wind speed promotes HAB growth while high wind speed disrupts growth. In 2011, there was weak wind which allowed the large amount of phosphorus from spring storms to sit in the western basin for longer than the average year (Kozacek, April 2014). In the literature review it was observed that machine learning forecasting studies often included

a lagged dependent variable as an independent variable. From this observation, a one month lagged CI was also considered as an input variable. The variables are analyzed in section 3.2 and were narrowed down to the nine independent and two dependent variables shown in Table 4 and 5 below.

Table 4. List of Final Independent Variables.

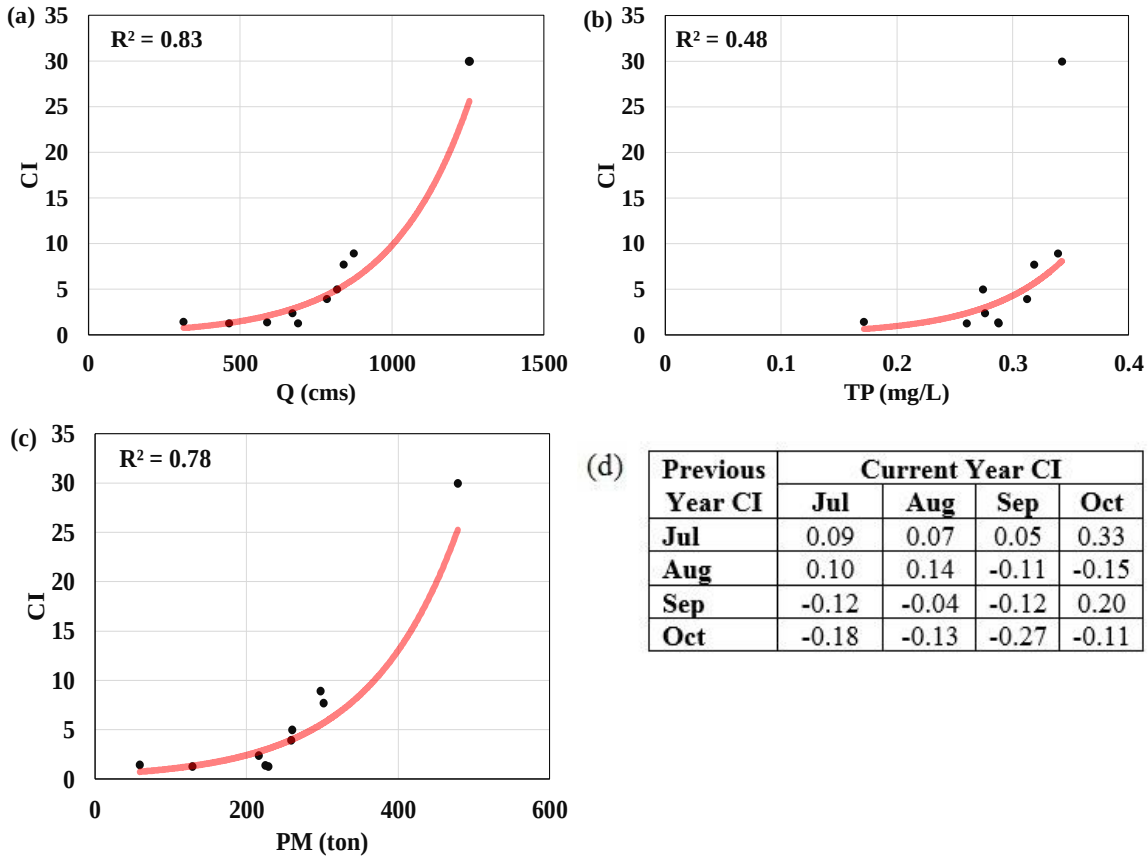
	Variable	Unit	Method	Source
Q	River Discharge	Cms	Average	NCWQR
TP	Phosphorus Concentration	mg/L	Average	NCWQR
PM	Phosphorus Mass	Ton	Average	NCWQR
SRP	Soluble Reactive Phosphorus	mg/L	Average	NCWQR
TKN	Total Kjeldahl Nitrogen	mg/L	Average	NCWQR
CI	Cyanobacterial Index	10^{20}	Average, Max	NOAA
Water	Water Temperature	C°	Average	USGS
Air	Air Temperature	C°	Average	USGS
Wind	Wind Speed	knots	Average	USGS

Table 5. List of Final Dependent Variables

Variable		Unit	Method	Source
CI	Cyanobacterial Index	10^{20}	Average	NOAA
CI	Cyanobacterial Index	10^{20}	Max	NOAA

3.2 Data Analysis and Variable Elimination

The first step of analyzing the data was looking at the correlations between the input and output variables. The first analysis performed was to determine the dependent variable to use in forecasting. The first set of plots (Figure 5) analyzed were Q, TP, and PM against CI for each year from 2002 to 2011. The three independent variables were all averaged for the accepted loading period from the literature review of March to June. The dependent variable, CI, for each year was the peak value from the four bloom months. The CI for the previous year was also analyzed against the CI for the current year (Figure 5).



(d)

Previous Year CI	Current Year CI			
	Jul	Aug	Sep	Oct
Jul	0.09	0.07	0.05	0.33
Aug	0.10	0.14	-0.11	-0.15
Sep	-0.12	-0.04	-0.12	0.20
Oct	-0.18	-0.13	-0.27	-0.11

Figure 5. Correlation of Peak CI and Nutrient Contributing Variables Averaged from the Nutrient Loading Period for 2002 to 2011: (a) Q Averaged from March to June vs. Peak CI, (b) TP Averaged from March to June vs. Peak CI, (c) PM Averaged from March to June vs. Peak CI, and (d) Previous Year vs. Current Year CI

The Q and CI relationship having a high correlation makes sense as often when there are more severe rain storms during the loading period the flow rate is high coupled with an increased phosphorus runoff. The correlation between the TP and CI is shown to not as great as Q and PM, with a value of 0.48 compared to 0.83 and 0.78. This is possibly explained as there can be times of high concentration of TP but low Q resulting in a lower amount of total phosphorus entering the Lake. The correlation of PM and CI is a combination of Q and TP. PM is the total amount of phosphorus entering the lake from the Maumee River. The correlation results between the current year CI and previous year show minimal correlation with the highest correlation coefficient of 0.33.

The second step in determining the optimal dependent variable to use in forecasting was looking at the correlations between the nutrient loading variables against Chl-a. The second set of plots (Figure 6) analyzed were Q, TP, and PM against Chl-a for each year from 2002 to 2011. The three independent variables were all averaged for the accepted loading period from the literature review of March to June. The dependent variable Chl-a for each year was the peak value from the four bloom months.

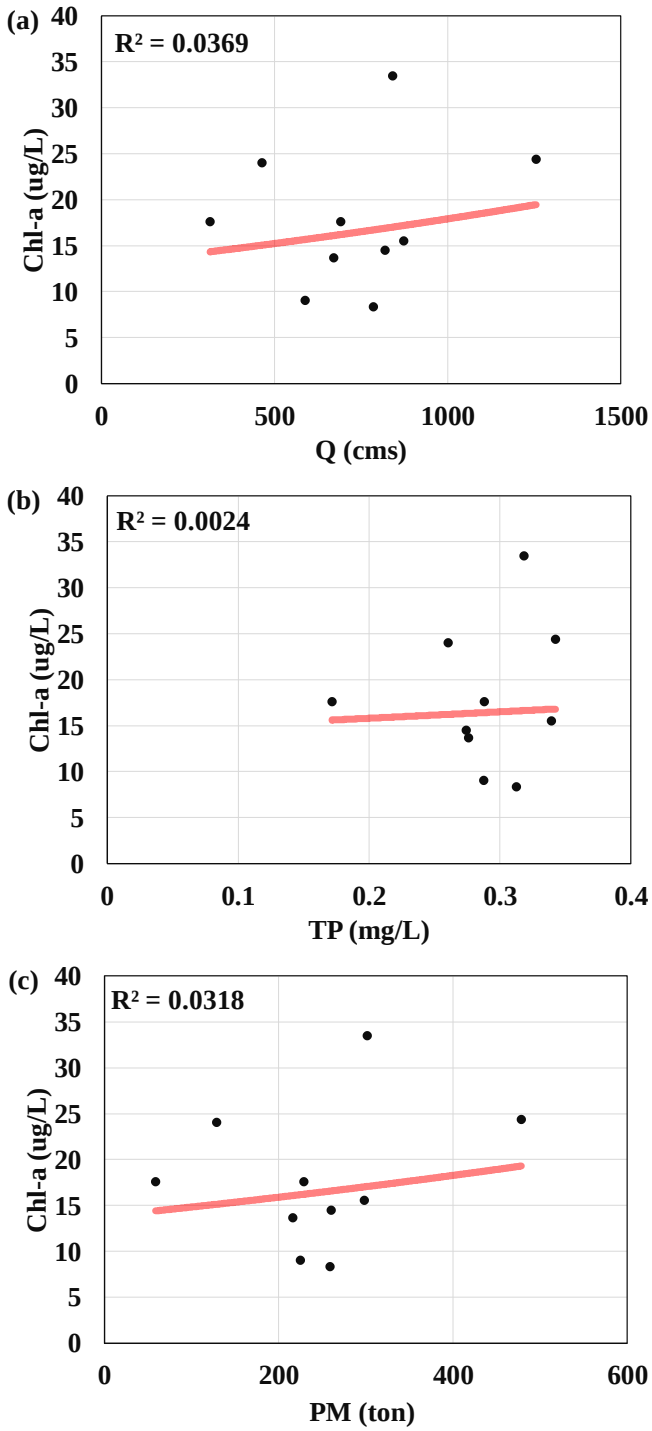


Figure 6. Correlation of Peak Chl-a and Nutrient Contributing Variables Averaged from the Bloom Period for 2002 to 2011: (a) Q Averaged from March to June vs. Peak Chl-a, (b) TP Averaged from March to June vs. Peak Chl-a, and (c) PM Averaged from March to June vs. Peak Chl-a

The correlations for all three variables Q, TP, and PM against the peak Chl-a values for the bloom months were all extremely low with a value less than 0.1 (Figure 6). Through the analysis of comparing CI and Chl-a against the nutrient contributing variables it was determined to only use CI as the dependent variables. After determining the optimal dependent variables, the next analysis looked at the importance of variables taken from the nutrient loading period. The set analyzed shows the averaged flow and phosphorus concentration from March to June (Figure 7).

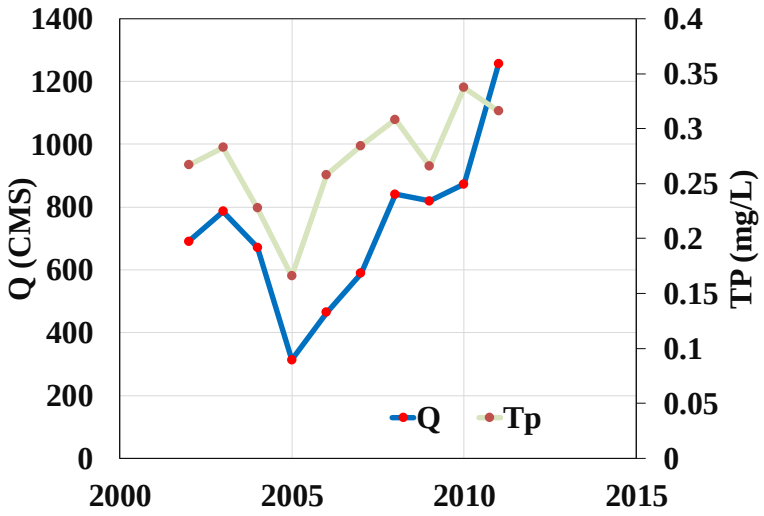


Figure 7. Comparison of Q and TP Averaged Each Year from March to June for 2002 to 2011

The analysis of Q and TP shows that they mostly follow the same trend which makes sense as the rain storms causes phosphorus runoff and would also cause an increase in flow rate. These two variables seem to follow a similar trend. However, when compared to CI separately, flow rate has a far better correlation (Figure 5). The next analysis of variables looked at the monthly distribution for the nutrient related variables (Figure 8).

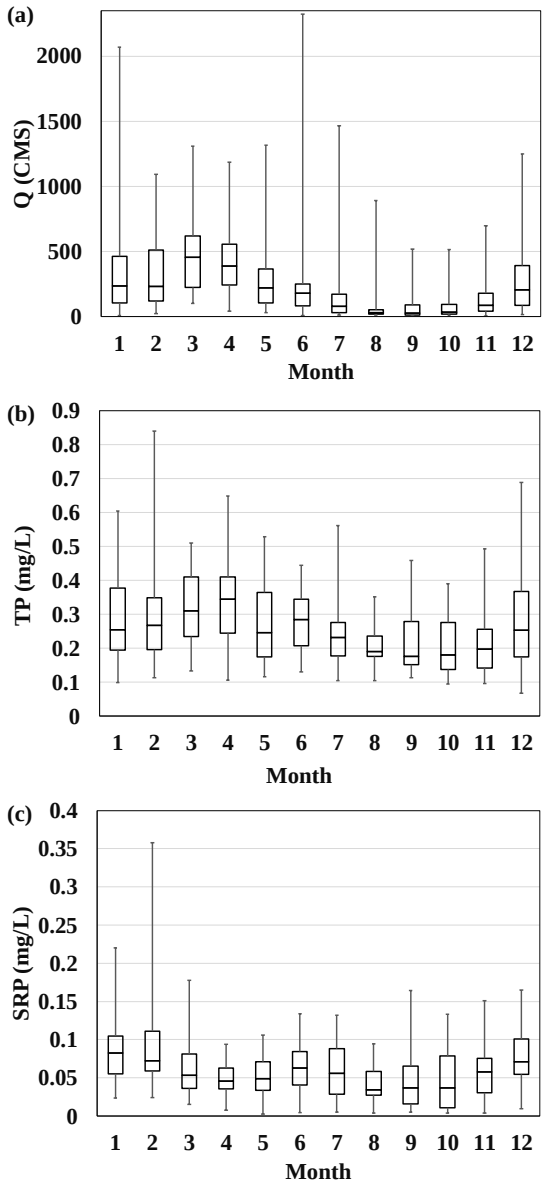


Figure 8. Monthly Distribution of Nutrient Related Variables from 1975 to 2015 (Heidelberg, 2017): (a) Q, (b) TP, and (c) SRP

The monthly distribution of nutrient related variables confirmed the accuracy of the nutrient loading months of March, April, May, and June. Both Q and TP values have their top two median values in March and April, the beginning of the nutrient loading period. During the bloom months July (7), August (8), September (9), and October (10) the median Q is often extremely low when compared to the nutrient loading period

resulting in a majority of the nutrients entering the lake before the blooms occur. The final two variables analyzed was averaged monthly flow values from the Maumee River and CI values from 2002 to 2007 shown in (Figure 9) below.

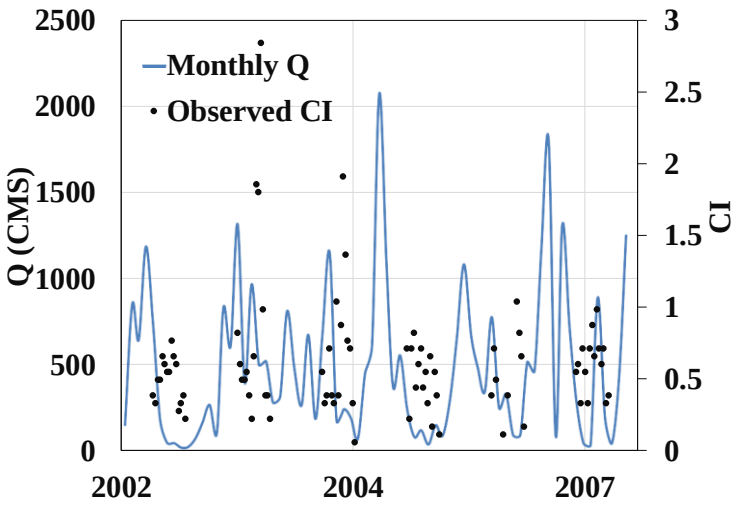


Figure 9. Observed CI vs Monthly Q from 2002 Through 2007

From observing Figure 9, it was confirmed that there was a disparity between times of peak flow and CI values. After confirming this disparity between values, determining the time to lag each variable was done in two methods. One method was through the literature review the accepted nutrient loading period of March to June. The second method to determine the averaging period and time to lag each variable was selected by analyzing the correlations between input variables and CI values using the Spearman rank correlation coefficient.

3.3 Models

3.3.1 Classification and Regression Tree

The CART training function used in this study is *fitctree* which follows a set of node splitting rules from (MathWorks, 2017). The first processes consist of computing the weighted impurity of each node *t*, i_t then approximates the chance that an observation is in the node (*t*) using Equation 1:

$$P(t) = \sum_{j \in T} w_j \quad (1)$$

where

w_j = Weight of the observation *j* with $w_j = 1/n$

n = Sample size

T = Set of all observation indices in node *t*

The next process is sorting all the predictors x_i, i = 1,…,p as splitting candidates or cut points. In the final step *fitctree* decides the optimal splits for each node (t) by maximizing the impurity gain ΔI (Equation 2) for all splitting candidates in x_i using the following process:

a. Splitting the observations in node t into left (t_L) and right (t_R) child nodes

b. Computing ΔI, for example looking at a splitting candidate, t_L and t_R and contains observation indices in sets T_L and T_R

$$\Delta I = P(T)i_t - P(T_L)i_{t_L} - P(T_R)i_{t_R} \quad (2)$$

The algorithm continues splitting branch nodes until one of the following occurs:

a. The set max number of splits is reached

b. A planned split results in the number of observations in a branch node to be fewer than the *MinParentSize*

c. The algorithm is unable to find a good split within a layer

In the final CART modeling, two model parameters were used: *MinParentSize* and *PredictorNames*. *MinParentSize* sets the minimum number of branch node observations and is used to control the decision tree depth. The default value of ten yields smaller trees with a low amount of observations. A value of one was set to create deep trees. The second parameter used was *PredictorNames* to give names to the variables in the decision tree corresponding to the x variables entered. After the final trees were made, the variable importance for each tree was determined using *predictorImportance*. *PredictorImportance* estimates the importance of each variable by summing the changes in the mean squared error from splits on each predictor and dividing by the sum of the number of branch nodes.

3.3.2 Artificial Neural Network

ANN models perform similar actions to the neurons in the human brain acquiring knowledge through a learning process that determines interneuron connection weights. They have three distinct layers (Figure 10). An input layer contains the known data in input nodes where the input data is rescaled to [-1,1]. A second a set of hidden layers with neurons with different sets of weights are determined from a training period with known inputs and outputs, and thirdly, the output layer which is determined by the hidden layer and is where the outputs are transformed to their original scale. The three layers are linked by a set of weights and biases determined by the learning algorithm. Figure 10 shows the general architecture of an ANN model.

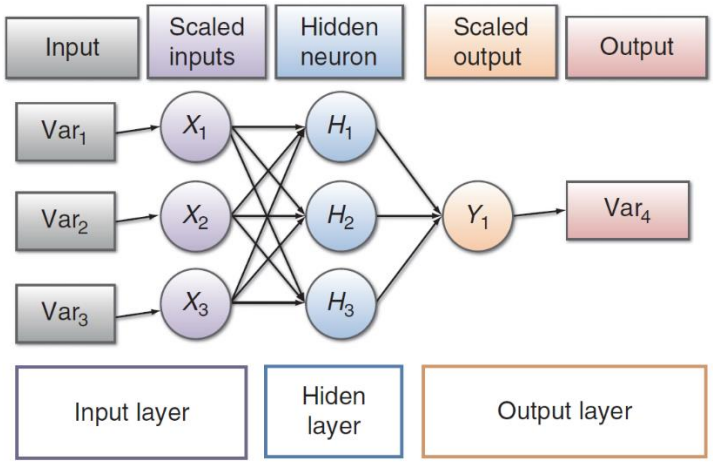


Figure 10. Example of a Simple ANN Network with Three Inputs, One Hidden Layer, Three Neurons, One Output Layer, and One Output Variable (Kang, 2011)

Neurons are the essential processing unit of ANN models and are connected to each other through links. Neurons consists of three phases: input, internal, and output. The neurons in the first hidden layer receive weighted signals from the input layer. The first phase is the input into the neuron and is the summing junction. The summing junction sums the weighted inputs with the following function (Equation 3):

$$U_n = \sum_{j=1}^{m} W_{nj} X_j \quad (3)$$

where

X_j = the j^{th} input signal from a total of m inputs

W_{nj} = the strength of connection weight from the j^{th} signal

U_n = the sum of the weighted inputs to neuron n

In the second phase, a bias B_n is added to the collective output U_n to determine the activation potential *Vn* of the neuron (Equation 4).

$$V_n = U_n + B_n \quad (4)$$

In the third phase, the activation potential is then passed to the transfer function φ (Equation 6) which computes the output Y_n of the neuron using the *tansig* neural transfer function (Equation 5).

$$\varphi = \frac{2}{1+\exp(-2*V_n)} - 1 \quad (5)$$

$$Y_n = \varphi\,(V_n) \quad (6)$$

The neurons pass information in a line through each of the hidden layers until the final hidden layer passes their output to the neurons in the output layer. The output of the neurons in the output layer is then rescaled back and is the output of the ANN model.

ANNs can learn and acquire knowledge about a problem through training. The training is the updating of weights and biases between the neurons. The ANN training function used in this study is the Bayesian regularization backpropagation (*trainbr*) with five hidden layers. The networking training function uses the Levenberg-Marquardt optimization to update weight and bias values that connect all the neurons and then determines the best combination to generate a network (MathWorks, 2017). The Levenberg-Marquardt algorithm used expresses the sum of squares of nonlinear functions by using an iterative technique to find the minimum of a function (Lourakis, 2005). The Bayesian regularization occurs in the Levenberg-Marquardt algorithm. The Jacobian jX is calculated with backpropagation with respect to the weight and bias variables X with each variable being adjusted by the Levenberg-Marquardt algorithm (MathWorks, 2017) shown in Equations 7, 8, and 9.

$$jj = jX * jX \quad (7)$$

$$je = jX * E \quad (8)$$

$$dX = -(jj+I*mu) / je \quad (9)$$

where

E = All errors

I = Identity matrix

mu = Adaptive value

The value of mu is adaptive and is increased by *mu_inc* until the change results in reduced performance and the change is made to the network and *mu* is decreased by *mu_dec*. Training continues until one of the following conditions occurs (Mathworks, 2017):

- The maximum number of repetitions is reached
- The maximum time allotted is reached
- Performance is minimized to the goal
- Performance gradient falls below the *min_grad*
- *mu* exceeds *mu_max*

Using the *trainbr* function makes the network hard to over train or overfit as it contains a criterion for stopping training and calculates effective weights and parameters (Burder, 2009). In the input and output layers, the data is transformed to a [-1,1] scale and transformed back to the original scale.

3.4 Model Input Variable Selection

Systematic selection for the optimal sets of input variables for machine learning models is important to improve the accuracy of the models. The final selection of input variables with the corresponding averaging periods and lag times was determined through two methods. In this study, two variable selection methods were used: method one and

two mentioned in the literature review. The first variable selection method, used in other Lake Erie HAB forecasting models, considers an accepted nutrient loading period with nutrient contributing variables from March to June. With this first variable selection method, an individual correlation analysis was performed to confirm viability of the nutrient loading period with the bloom months (Equation 10).

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (10)$$

where

r = Pearson product moment correlation coefficient

$\bar{x}$ = Average of the independent variable array

$\bar{y}$ = Average of the dependent variable array

The second variable selection method considers the nutrient loading variables as well as climate variables and uses the Spearman rank correlation coefficient method to select averaging and lag periods for each month. The Spearman method calculates ρ using (Equation 11) and then transforms ρ into a *p*-value using exact permutation distributions and *p*-values less than 0.05 represent high significance often in statistical analyses. A *p*-value of 0.05 represents that the corresponding input variable is statistically significant with 95% confidence. In this method, two hundred and twenty-four different combinations of averaging and lag periods were considered and *p*-values calculated for the independent variables separately for each bloom month. The Spearman rank correlation first ranks the independent variable x and dependent variable y and then calculates ρ (Equation 11).

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)} \quad (11)$$

where

ρ = Spearman rank correlation coefficient

d_i = Difference in ranks between corresponding x and y variables

n = Total number of values in the data set

Two training periods were considered for each method: 2002 to 2011 and 2002 to 2013. The first method was the nutrient loading period determined by the literature review. To confirm that the nutrient loading period selection is viable for forecasting HABs, an individual correlation analysis was performed on numerous averaging periods and lag times. The second method for determining the correct set of input variables was using Spearman's rank correlation coefficient. A total of two hundred and twenty-four variables were considered using different lag times and averaging periods for nine variables.

3.4.1 Nutrient Loading Period Selection

The nutrient loading period for the nutrients for HABs was determined to be March, April, May, and June through the literature review. An individual correlation analysis with up to thirty different averaging periods and time lags considered for each variable (Table 6) was performed on the main nutrient phosphorus for the HABs in Lake Erie and the flow from the Maumee River which is the main source for the phosphorus to enter the lake (Table 7). The individual correlations were also looked at for air temperature and wind speed as they are also important variables that can limit or promote growth of HABs (Table 7).

Table 6. Number of Averaging Periods and Time Lags Considered for the Loading Period Correlation Analysis.

	Averaging Periods	Time Lags	Total
Flow	6	5	30
PM	6	5	30
TP	6	5	30
Air	4	2	7
Wind	4	1	4
SRP	6	5	30
CI	1	1	1

Table 7. Top Two Individual Correlations Comparing Various Averaging Periods and Time Lags With Observed CI for the Loading Period Analysis.

		July		August		September		October	
		1	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Q	**Correlation**	0.84	0.81	0.95	0.89	0.87	0.87	0.84	0.79
	Method	2avg	3avg	5avg	6avg	3avg	3avg	2avg	3avg
	Lag	t-2	t-1	t-2	t-2	t-4	t-5	t-5	t-4
PM	**Correlation**	0.79	0.78	0.96	0.93	0.83	0.83	0.77	0.76
	Method	3avg	4avg	5avg	4avg	4avg	3avg	2avg	4avg
	Lag	t-2	t-1	t-2	t-3	t-3	t-4	t-5	t-4
TP	**Correlation**	0.57	0.55	0.69	0.68	0.65	0.60	0.60	0.55
	Method	2avg	4avg	5avg	4avg	4avg	2avg	4avg	2avg
	Lag	t-2	t-2	t-2	t-3	t-4	t-4	t-5	t-5
Air	**Correlation**	0.65	0.60	0.27	0.23	0.57	0.56	0.32	0.31
	Method	1avg	3avg	3avg	2avg	2avg	4avg	4avg	3avg
	Lag	t	t	t-1	t-1	t-1	t-1	t-1	t-1
Wind	**Correlation**	0.45	0.30	0.19	0.16	0.36	0.33	-0.25	-0.28
	Method	4avg	2avg	3avg	4avg	4avg	1avg	2avg	3avg
	Lag	t	t	t	t	t	t	t	t
SRP	**Correlation**	0.75	0.61	0.53	0.46	0.49	0.32	0.34	0.32
	Method	2avg	4avg	1avg	3avg	5avg	4avg	2avg	2avg
	Lag	t-5	t-3	t-1	t-4	t-2	t-3	t-1	t-2
CI	**Correlation**			0.92		0.85	0.75	0.61	0.49
	Method			1avg		1avg	1avg	1avg	1avg
	Lag			t-1		t-1	t-2	t-1	t-2

The individual correlations for flow and phosphorus were higher than the air temperature and wind speed as expected. The amount of nutrients entering the lake have a large effect on the size of the blooms. The wind speed and air temperature are able to

promote or discourage the growth of HABs however they are unable to cause or stop blooms individually. The top correlations for flow, phosphorus concentration, and total mass of phosphorus were around the nutrient period. Through the analysis of individual correlations, it was decided to use all nutrient contributing variables for CART modeling and previous CI as well for ANN modeling for the first method of forecasting shown in Table 8.

Table 8. Final Loading Period Inputs for ANN and CART Models with the Addition of Previous Month CI for ANN Models for Both Training Periods.

		July	**August**	**September**	**October**
Q	**Method**	4avg	4avg	4avg	4avg
	Lag	t-1	t-2	t-3	t-4
TP	**Method**	4avg	4avg	4avg	4avg
	Lag	t-1	t-2	t-3	t-4
PM	**Method**	4avg	4avg	4avg	4avg
	Lag	t-1	t-2	t-3	t-4
SRP	**Method**	4avg	4avg	4avg	4avg
	Lag	t-1	t-2	t-3	t-4
CI	**Method**		1avg	1avg	1avg
	Lag		t-1	t-1	t-1

3.4.2 Spearman Selection

A Spearman rank correlation coefficient analysis was performed in the process of determining the optimal set of each input variables for each month. The first step was creating an ANN model using all two hundred and twenty-four variables using 100% training to determine if using all considered variables is a viable forecasting input. The results for September are shown in Figure 11.

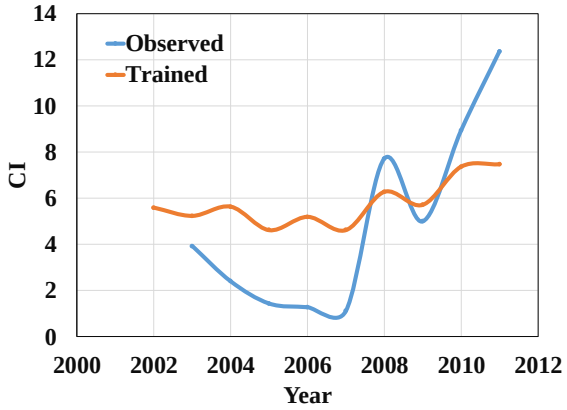


Figure 11. Comparison of Observed CI and Trained CI for September Using All Two Hundred and Twenty-Four Variables

After reviewing the result, it was determined that to select the optimal set of input variables the number of variables had to be reduced. When using all considered lag times and averaging periods, the model was overwhelmed by too many prediction variables while not correlating between input and output and unable to make viable forecasts.

The second step was running the Spearman algorithm with the two hundred and twenty-four averaging periods and lag times for the bloom months of July, August, September, and October and selecting the variables with a *p*-value less than 0.05 for both training periods. The result of this step reduced the total number of variables to 31, 34, 19, and 18 for the 2002 to 2011 training period and 37, 15, 17, and 6 for the 2002 to 2013 training period for July, August, September, and October, respectively.

The third step was the removal of significantly overlapped variables in order to reduce bias in the model. Variables that were overlapped by two-thirds by a variable retained were removed from the final set of inputs. For example, if Q from February to June was selected from step 2 it would be removed if Q from January to June was also selected from step 2. The result of this step reduced the total number of variables to 15,

14, 11, and 12 for the 2002 to 2011 training period and 16, 9, 9, and 6 for the 2002 to 2013 training period for July, August, September, and October, respectively.

The final selection of input variables from the Spearman rank correlation analysis are shown in Table 9. For example, for July Q1,6 represents a 1-month lag with a 6-month averaging period averaging the flow rate from January to June.

Table 9. Final Spearman Selected Inputs for ANN and CART Models with the Addition of Previous Month CI for ANN Models for Both Training Periods.

July		**August**		**September**		**October**	
(02-11)	**(02-13)**	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**
Q5,1	Q3,1	Q5,1	Q5,1	Q3,1	Q3,1	Q5,2	Q4,1
Q4,2	Q4,1	Q4,2	Q4,2	Q6,1	Q6,1	Q1,6	PM2,5
Q3,3	Q6,1	Q1,6	Q1,6	Q3,3	Q3,3	TP1,6	TKN4,2
Q1,6	Q2,2	PM5,1	PM5,1	Q1,6	Q1,6	PM5,2	Wind4,1
TP1,5	Q3,2	PM1,6	PM4,2	TP3,4	PM3,1	PM1,6	Wind4,2
PM3,1	Q1,5	SRP6,1	PM1,6	PM3,1	PM3,3	TKN1,2	CI1,1
PM4,1	TP3,1	SRP5,2	SRP5,2	PM3,3	PM1,6	TKN4,2	
PM6,1	TP1,5	TKN1,6	Water2,1	PM1,6	TKN3,4	Air3,2	
PM2,2	PM3,1	Water2,1	CI1,1	TKN1,1	CI1,1	Air3,4	
PM1,5	PM2,2	Water1,2		TKN3,4		Wind3,2	
TKN5,2	PM3,2	Water2,2		CI1,1		Wind1,4	
TKM3,3	PM1,5	Water1,3				CI1,1	
TKM4,3	TKN5,2	Water1,6					
TKM1,6	TKN2,3	CI1,1					
Water3,4	TKN1,6						
	Water3,4						

CHAPTER IV

RESULTS AND DISCUSSION

This chapter first details the different CART models created. This is followed by examining the accuracy of the CART models. Lastly, this chapter goes through the different ANN models detailed followed by an accuracy analysis.

4.1 Classification and Regression Tree

Four groups of CART models were made: two for each variable selection method and two training periods. Two class systems are considered, a five-class system based on the breakdown for the peak bloom forecast discussed in the literature review and a simplified three class system. The first classification was made for five classes (see 2.2.2), Class 1 (CI<2), Class 2 (CI>=2 to CI <4), Class 3 (CI>=4 to CI<7), Class 4 (CI>=7 to CI<10), and Class 5 (CI>=10). The second classification was made for three classes, Class 1 (CI<2), Class 2 (CI>=2 to CI <=7), and Class 3 (CI>7). The CART models were trained using two methods, data from 2002 to 2011 (Train(02-11)) and 2002 to 2013 (Train(02-13)). First, CART models were created for the two classification systems to determine the

optimal class breakdown. The CART decision tree for Spearman five classes for October is shown in Figure 12 below.

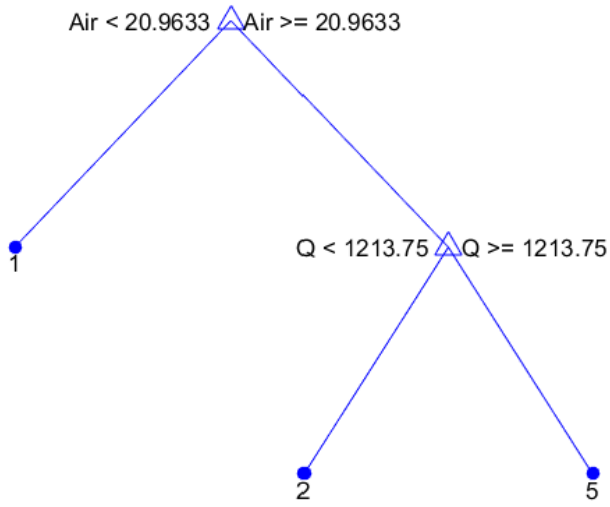


Figure 12. October Spearman Five Class Decision Tree Using Train(02-11)

The results from the five-class decision tree showed a three-class gap in the prediction (Figure 12). During the training period of 2002 to 2011, the CI values for October had a peak value of nearly thirty while the second peak for October was only three. The year of 2011 was the first appearance of massive severe HABs. Before 2011, the most severe bloom since 2002 was nine. The 2011 HAB was also the only year on the CI record (2002 to 2015) where the HAB peaked in October and not August or September. Five-class decision trees were also made for the remaining bloom months which also resulted in a three-class gap for the August prediction. The CART trees for the loading period variable selection method are shown first below (Figure 13).

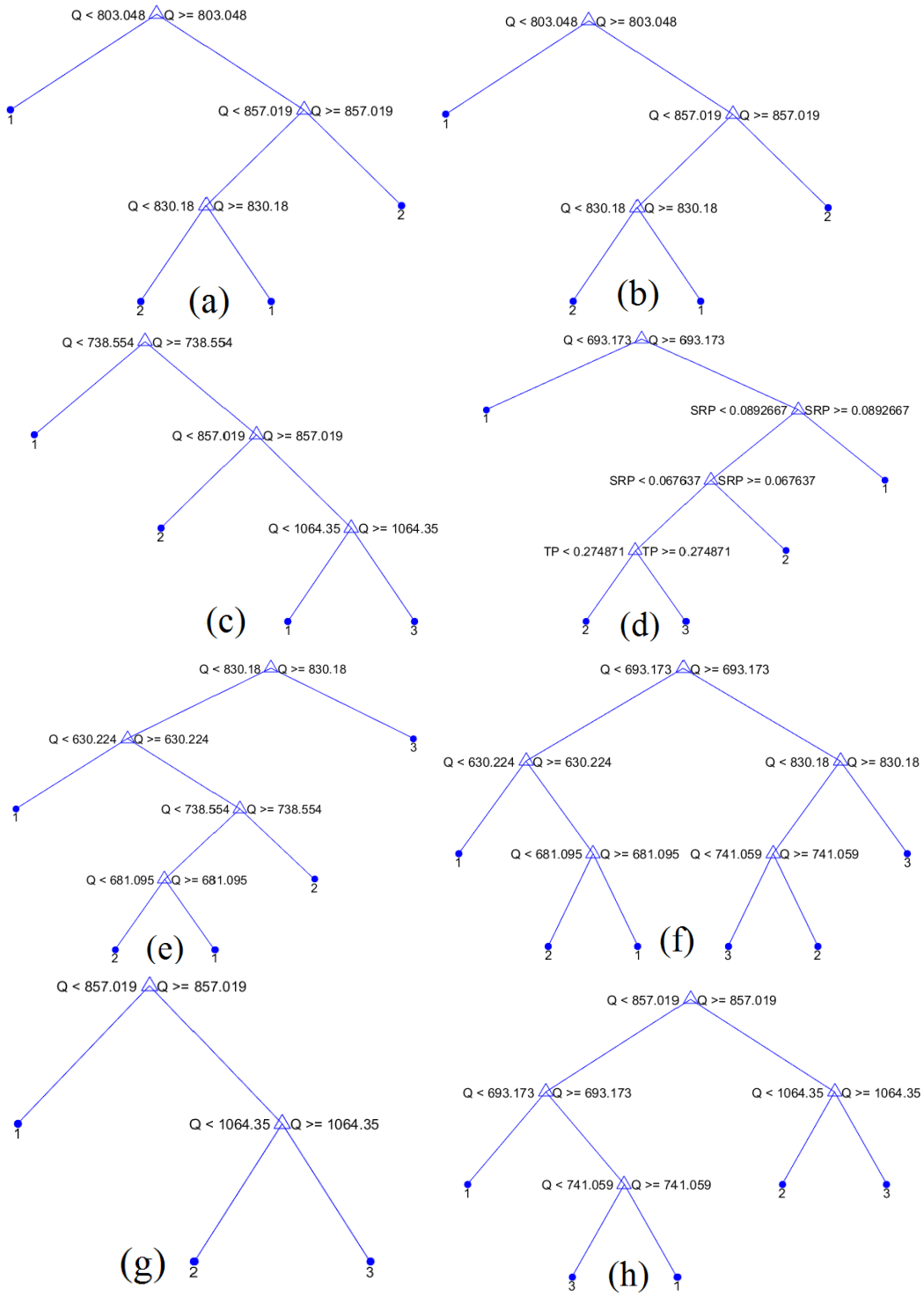


Figure 13. Loading Period Three Class Monthly Decision Trees: (a) July Train(02-11), (b) July Train(02-13), (c) August Train(02-11), (d) August Train(02-13), (e) September Train(02-11), (f) September Train(02-13), (g) October Train(02-11), and (h) October Train(02-13)

The decision trees for August, September, and October for both training methods are able to forecast all three classifications. The decision trees for July are only able to forecast Class 1 or 2 with the two training periods because from 2002 to 2013 there were no blooms in July that are classified as Class 3 bloom. However, in 2015, there was Class 3 bloom in July. When the CART model for July is trained using the full 2002 to 2015 data set, the new decision tree is able to forecast Class 3 blooms. Interestingly, for July, September, and October only Q was used in the decision trees where August, for the longer training period, starts to use SRP and TP. The eight decision trees made for the loading period selection method and both training periods were next analyzied for their precision for predicting the monthly HABs with predicted classes after training period shown in parentheses (Table 10 and 11).

Table 10. Loading Period Three Classes Result Matrix for Train(02-11) and Predicted(12-15).

	Predicted					
	Class	**1**	**2**	**3**	**Sum**	**Precision**
Observed (July)	**1**	10 (3)	0	0	10	100.00% (75%)
	2	0	3	0	3	100.00%
	3	0	1 (1)	0	1	0.00% (0%)
	Sum	10	4	0	14	**92.86% (75%)**
Observed (August)	**1**	7 (1)	0	0	7	100.00% (100%)
	2	0	4 (1)	0	4	100.00% (100%)
	3	2 (2)	0	1	3	33.33% (0%)
	Sum	9	4	1	14	**85.71% (50%)**
Observed (September)	**1**	5 (1)	0	0	5	100.00% (100%)
	2	0	3	1 (1)	4	75.00% (0%)
	3	1 (1)	0	4 (1)	5	80.00% (50%)
	Sum	6	3	5	14	**85.71% (50%)**
Observed (October)	**1**	10 (2)	0	0	10	100.00% (100%)
	2	0	1	0	1	100.00%
	3	1 (1)	1 (1)	1	3	33.33% (0%)
	Sum	11	2	1	14	**85.71% (50%)**

NOTE: ‘Underlined’ indicates overall precision for that particular month and period.

Table 11. Loading Period Three Classes Result Matrix for Train(02-13) and Predicted(14-15).

	Predicted					
	Class	**1**	**2**	**3**	**Sum**	**Precision**
Observed (July)	**1**	10 (1)	0	0	10	100.00% (100%)
	2	0	3	0	3	100.00%
	3	0	1 (1)	0	1	0.00%(0%)
	Sum	10	4	0	14	**92.86% (50%)**
Observed (August)	**1**	7	0	0	7	100.00%
	2	0	4 (1)	0	4	100.00% (100%)
	3	0	1 (1)	2	3	66.67%(0%)
	Sum	7	5	2	14	**92.86% (50%)**
Observed (September)	**1**	5	0	0	5	100.00%
	2	0	3	1 (1)	4	75.00% (0%)
	3	0	0	5 (1)	5	100.00% (100%)
	Sum	5	3	6	14	**92.86% (50%)**
Observed (October)	**1**	10 (1)	0	0	10	100.00% (100%)
	2	0	1	0	1	100.00%
	3	0	1 (1)	2	3	66.67% (0%)
	Sum	10	2	2	14	**92.86% (50%)**

NOTE: 'Underlined' indicates overall precision for that particular month and period.

The overall precision for both training periods is high with the lowest being 85.7% during the Train(02-11) method. The precision for the Train(02-11) is high when predicting Class 1 or 2 blooms however it is low when predicting Class 3 blooms ranging from 33.3% to 80% for August, September, and October with the highest accuracy in September. However, when increasing the training period to Train(02-13), the precision range increases from 66.7% to 100% except for July where the decision tree is still unable to forecast Class 3 blooms. In the early years of recorded CI values, the blooms were small in comparison to the recent blooms which results in there being a small amount of data for CART to forecast the high-class blooms in some months. This is shown in the result matrixes as the predictions have a higher precision for forecasting the low-class blooms. When extending the training period to include the larger blooms of

recent years, the overall precision increases from 87.5% to 92.9%. Next the CART decision trees for the Spearman selection method are shown (Figure 14).

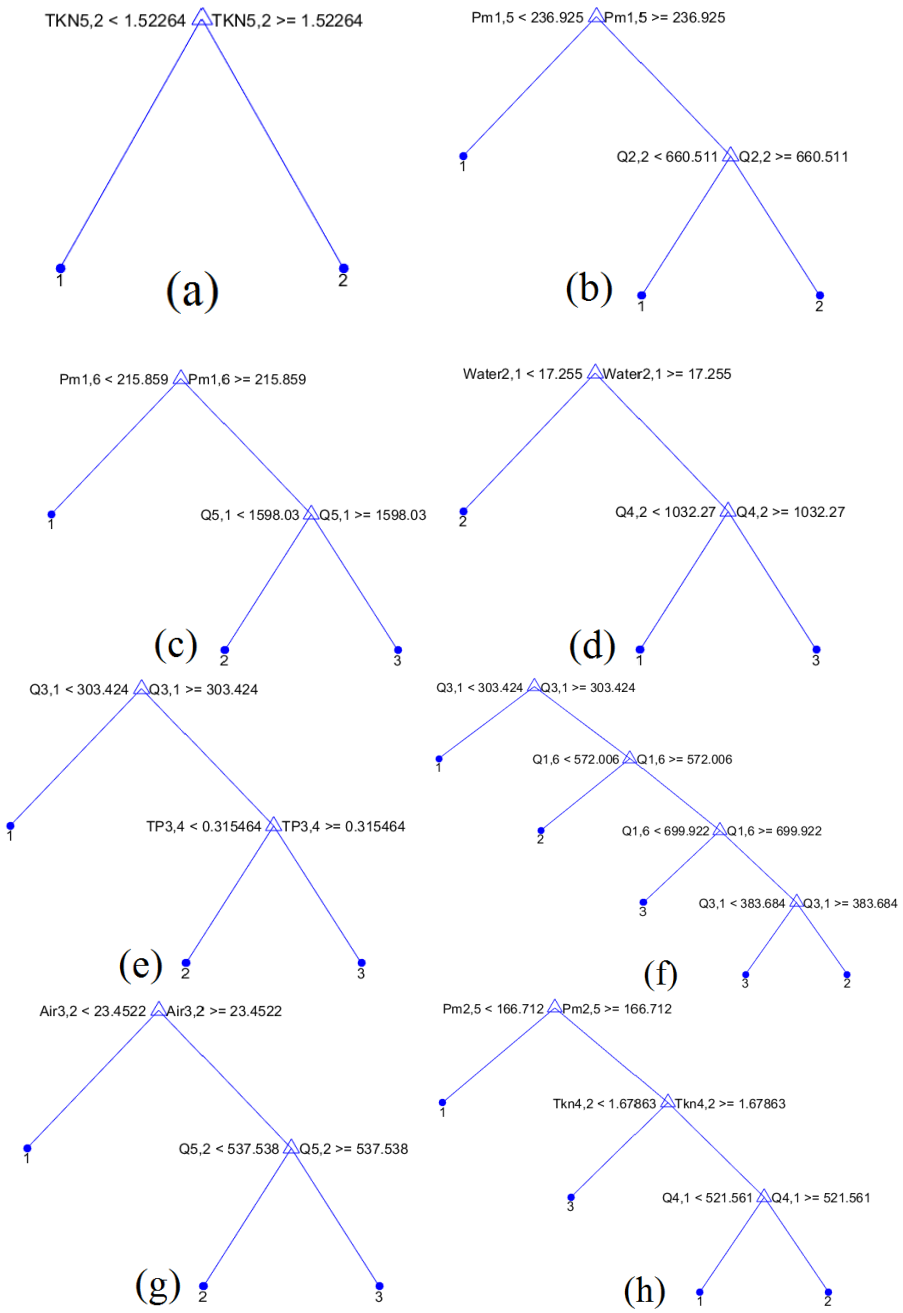


Figure 14. Spearman Selected Three Class Monthly Decision Trees: (a) July Train(02-11), (b) July Train(02-13), (c) August Train(02-11), (d) August Train(02-13), (e) September Train(02-11), (f) September Train(02-13), (g) October Train(02-11), and (h) October Train(02-13)

The decision trees for August, September, and October for both training methods are able to forecast all three classifications. Similar to the loading period selection method, decision trees for July are only able to forecast Class 1 or 2 with both training periods as from 2002 to 2013 there were no blooms in July that were classified as Class 3 bloom. Also, similar to the loading period method when modeled using the full 2002 to 2015 data set to train, the new decision tree changes are able to forecast Class 3 blooms. Unlike the loading period decision trees, the trees for the Spearman selection method uses a variety of variables for the trees. The eight decision trees made for the Spearman selection method and both training periods were next analyzied for their precision with predicted classes after training period shown in parentheses (Table 12 and 13).

Table 12. Spearman Three Classes Result Matrix for Train(02-11) and Predicted(12-15).

	Predicted					
	Class	**1**	**2**	**3**	**Sum**	**Precision**
Observed (July)	**1**	9 (2)	1 (1)	0	10	90.00% (66.7%)
	2	0	3	0	3	100.00%
	3	0	1 (1)	0	1	0.00% (0%)
	Sum	9	5	0	14	**85.71% (50%)**
Observed (August)	**1**	7 (1)	0	0	7	100.00% (100%)
	2	1 (1)	3	0	4	75.00% (0%)
	3	1 (1)	1 (1)	1	3	33.33% (0%)
	Sum	9	4	1	14	**78.57% (25%)**
Observed (September)	**1**	5 (1)	0	0	5	100.00% (100%)
	2	0	4 (1)	0	4	100.00% (100%)
	3	0	2 (2)	3	5	60.00% (0%)
	Sum	5	6	3	14	**85.71% (50%)**
Observed (October)	**1**	9 (1)	1 (1)	0	10	90.00% (50%)
	2	0	1	0	1	100.00%
	3	2 (2)	0	1	3	33.33% (0%)
	Sum	11	2	1	14	**78.57% (25%)**

NOTE: 'Underlined' indicates overall precision for that particular month and period.

Table 13. Spearman Three Classes Result Matrix for Train(02-13) and Predicted(14-15).

	Predicted					
	Class	**1**	**2**	**3**	**Sum**	**Precision**
Observed (July)	**1**	10 (1)	0	0	10	100.00% (100%)
	2	0	3	0	3	100.00%
	3	1 (1)	0	0	1	0.00% (0%)
	Sum	11	3	0	14	**92.86% (50%)**
Observed (August)	**1**	7	0	0	7	100.00%
	2	0	3	1 (1)	4	75.00% (0%)
	3	0	1(1)	2	3	66.67% (0%)
	Sum	7	4	3	14	**85.71% (0%)**
Observed (September)	**1**	5	0	0	5	100.00%
	2	0	3	1 (1)	4	75.00% (0%)
	3	0	1 (1)	4	5	80.00% (0%)
	Sum	5	4	5	14	**85.71% (0%)**
Observed (October)	**1**	10 (1)	0	0	10	100.00% (100%)
	2	0	1	0	1	100.00%
	3	0	0	3 (1)	3	100.00% (100%)
	Sum	10	1	3	14	**100.00% (100%)**

NOTE: 'Underlined' indicates overall precision for that particular month and period.

The overall precision for both training periods is average with the lowest being 78.6% during the Train(02-11) method. Similar to the loading period method, the precision for the Train(02-11) is high when predicting Class 1 or 2 blooms however it is low when predicting Class 3 blooms ranging from 33.3% to 60% for August, September, and October with the highest accuracy in September.

During the first training period, August and October both only had one Class 3 bloom whereas September had three. However, when increasing the training period to Train(02-13), the overall precision for all of the trees immensely increases as the longer training period with 2013 bloom gave the August, September, and October models another Class 3 bloom to be used in training. The precision for Class 3 blooms increased from 33%, 60%, and 33% to 67%, 80%, and 100% for August, September, and October, respectively, with the extended training period. After the extended training period for

August and September, the predictions for 2014 and 2015 showed opposite predictions: predicting Class 2 for Class 3 blooms and Class 3 for Class 2 blooms, respectively. October showed the greatest improvement in precision from 78.6% to 100% when extending the training period. In the early years of recorded CI values, the blooms were small in comparison to the recent blooms which results in there being a small amount of data for CART to forecast the large blooms. When extending the training period to include the larger blooms of recent years, the overall precision increases from 82.1% to 91%. Next, the variable importance for both selection methods and training periods were analyzed (Table 14, 15 and 16).

Table 14. Loading CART Variable Importance for Both Training Periods.

	July		**August**		**September**		**October**	
Variable	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**
Q	100%	100%	100%	46%	100%	100%	100%	100%
TP	0%	0%	0%	20%	0%	0%	0%	0%
PM	0%	0%	0%	0%	0%	0%	0%	0%
SRP	0%	0%	0%	34%	0%	0%	0%	0%

Table 15. Spearman CART Variable Importance for Both Training Periods for July and August.

July				August			
Train(02-11)		**Train(02-13)**		**Train(02-11)**		**Train(02-13)**	
Q5,1	0%	Q3,1	0%	Q5,1	28%	Q5,1	0%
Q4,2	0%	Q4,1	0%	Q4,2	0%	Q4,2	46%
Q3,3	0%	Q6,1	33%	Q1,6	0%	Q1,6	0%
Q1,6	0%	Q2,2	0%	PM5,1	0%	PM5,1	0%
TP1,5	0%	Q3,2	0%	PM1,6	72%	PM4,2	0%
PM3,1	0%	Q1,5	0%	SRP6,1	0%	PM1,6	0%
PM4,1	0%	TP3,1	0%	SRP5,2	0%	SRP5,2	0%
PM6,1	0%	TP1,5	0%	TKN1,6	0%	Water2,1	54%
PM2,2	0%	PM3,1	0%	Water2,1	0%		
PM1,5	0%	PM2,2	0%	Water1,2	0%		
TKN5,2	100%	PM3,2	0%	Water2,2	0%		
TKM3,3	0%	PM1,5	67%	Water1,3	0%		
TKM4,3	0%	TKN5,2	0%	Water1,6	0%		
TKM1,6	0%	TKN2,3	0%				
Water3,4	0%	TKN1,6	0%				
		Water3,4	0%				

Table 16. Spearman CART Variable Importance for Both Training Periods for September and October.

September				October			
(02-11)		**(02-13)**		**(02-11)**		**(02-13)**	
Q3,1	55%	Q3,1	69%	Q5,2	29%	Q4,1	21%
Q6,1	0%	Q6,1	0%	Q1,6	0%	PM2,5	48%
Q3,3	0%	Q3,3	0%	TP1,6	0%	TKN4,2	31%
Q1,6	0%	Q1,6	31%	PM5,2	0%	Wind4,1	0%
TP3,4	45%	PM3,1	0%	PM1,6	0%	Wind4,2	0%
PM3,1	0%	PM3,3	0%	TKN1,2	71%		
PM3,3	0%	PM1,6	0%	TKN4,2	0%		
PM1,6	0%	TKN3,4	0%	Air3,2	0%		
TKN1,1	0%			Air3,4	0%		
TKN3,4	0%			Wind3,2	0%		
				Wind1,4	0%		

For the loading period decision trees, interestingly, only Q was used for both training periods for all months except for August Train(02-13). When comparing the individual correlations from Table 7 for the loading period selection analysis for all months except August, Q had the highest value. For both selection methods Q is in every decision tree except for July Train(02-11) where only TKN is used in the Spearman method. In the

individual correlation analysis in Table 7, PM had the second highest correlation behind Q which is shown in the variable importance for both methods being selected the second most often in the decision trees.

4.2 Artificial Neural Network

Four groups of ANN models were made for each bloom month: two for each variable selection method and two for separate time periods. The variable selection methods were from the nutrient loading period and Spearman selected. The two-time periods are Train(02-11) and Train(02-13) as well as two predicting periods to be from 2012 to 2015 (Predicted(12-15)) and 2014 to 2015 (Predicted(14-15)). The two-separate time periods were combined into the same figure resulting in eight figures below showing the results for the two selection methods for each month starting with July (Figure 15).

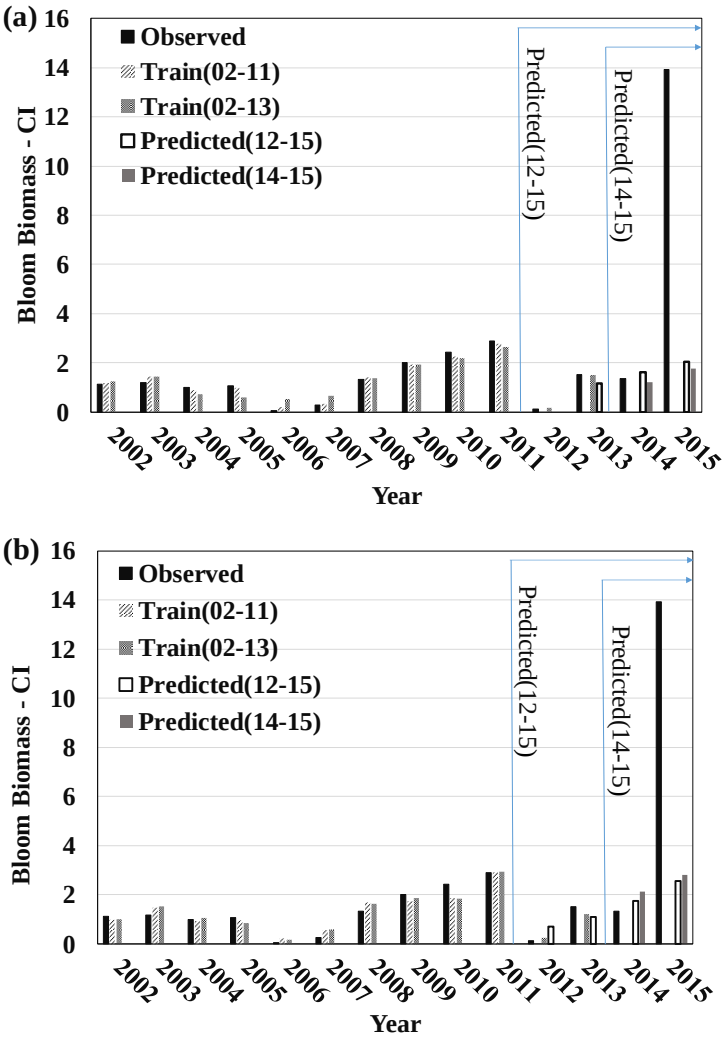


Figure 15. ANN Results for Both Selection Methods and Averaging Periods: (a) Spearman July and (b) Loading Period July

The predictions for July for both selection methods and training periods overall is good except for 2015. For both methods increasing the training period had little effect on the final predictions as 2012 and 2013 did not have any special cases of blooms to increase the accuracy of forecasts in July. Before 2015, the July peak CI value was nearly three however in 2015 there was a bloom of nearly fourteen resulting in the ANN model to have never trained for a bloom of that magnitude for the month of July. As the training period extends it will be possible for the ANN models to more closely predict the high

magnitude blooms in July. The results for the two selection methods for August is shown below (Figure 16).

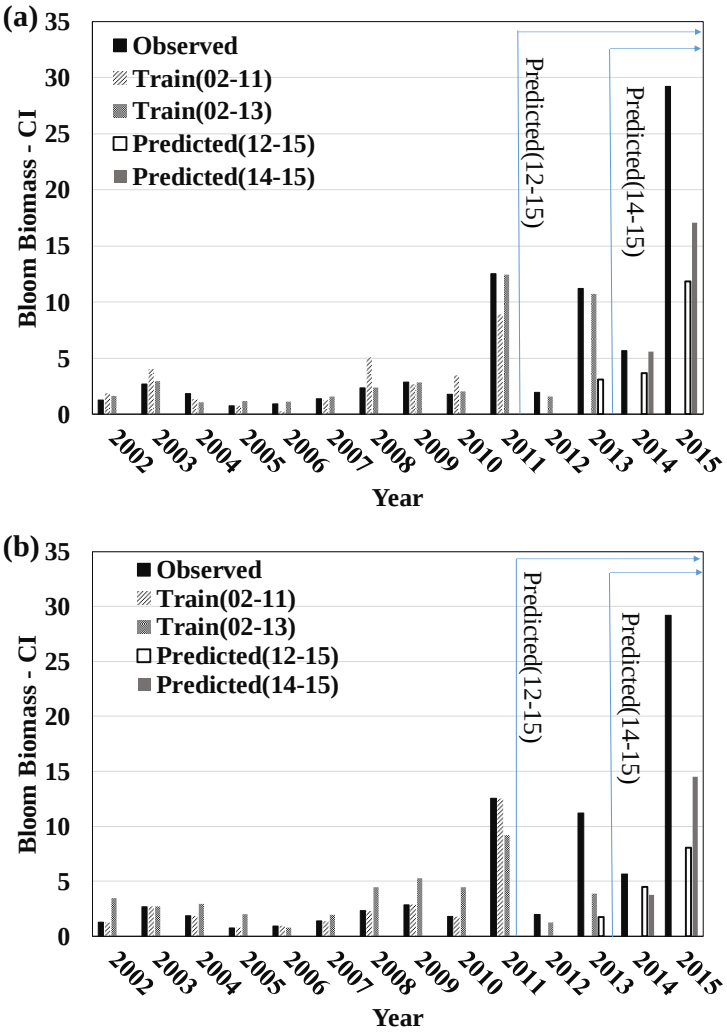


Figure 16. ANN Results for Both Selection Methods and Averaging Periods: (a) Spearman August and (b) Loading Period August

The predictions for August showed varying accuracy for both prediction methods for each year with both selection methods having similar predictions for each year. Both methods had a problem predicting the 2013 bloom. From 2002 to 2011, there was only

one bloom with a high magnitude, in 2011, resulting in the ANN models not being able to train more than one of the high magnitude of blooms.

The 2015 bloom was a special case with an unusual loading period with the most nutrients entering the lake in June and July as well as the magnitude being more than double previously recorded in August. Both methods under predicted the 2015 bloom however they still both predicted a large magnitude bloom over the previously recorded max CI value for both extended training periods.

For both methods, extending the training period improved the prediction for 2015 with a slightly larger increase in accuracy for the loading period method. Both methods showed that it is possible for the ANN models to forecast blooms higher than the magnitude they were trained with when predicting the 2015 bloom. The results for the two selection methods for September is shown below (Figure 17).

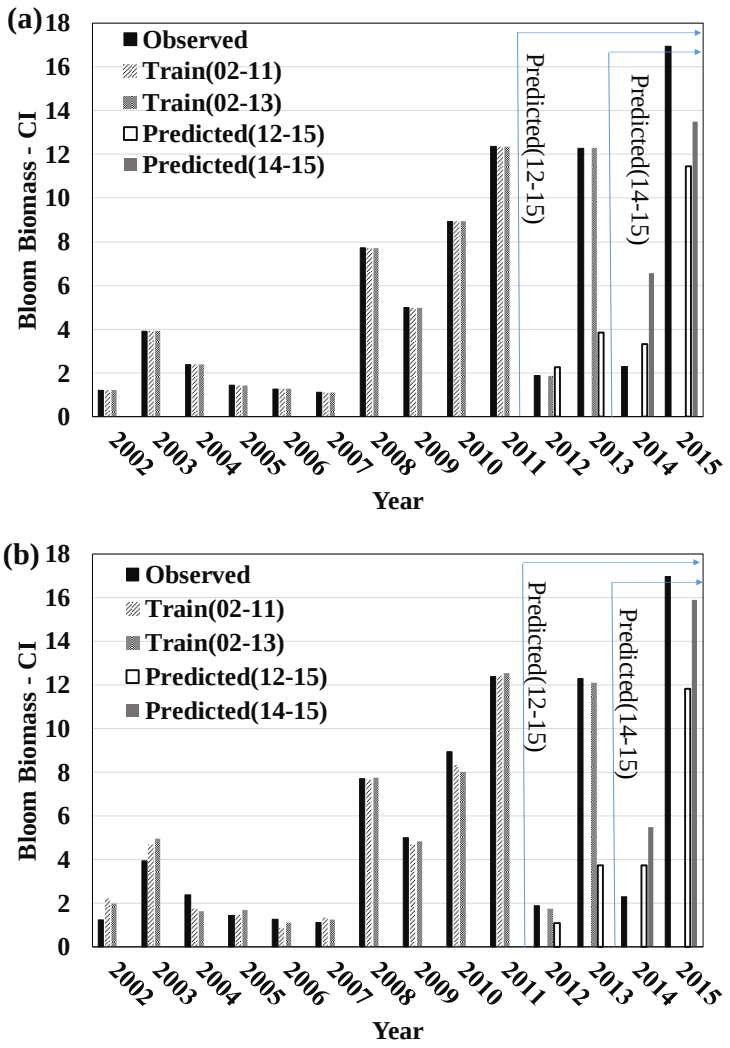


Figure 17. ANN Results for Both Selection Methods and Averaging Periods: (a) Spearman September and (b) Loading Period September

The predictions for September showed decent results for all predicted years for both methods except 2013 where both methods vastly under predict the 2013 bloom in September. Both methods saw an improvement in their 2015 prediction by increasing the training period with the loading period method showing a larger improvement with the increased training. The correlation coefficient for the Spearman method increased from 0.87 to 0.96 while the loading period method increased from 0.86 to 0.98 when extending

the training period. Interestingly, both methods overpredicted the 2014 bloom and the increased training caused both methods to overpredict even more.

The overall predictions for September are fairly accurate with correlation coefficients of 0.96 for the Spearman method and 0.98 for the loading period method with the extended training period and is possibly explained by the history of September. September has a history of higher magnitude blooms allowing the ANN models to have increased training for the larger blooms compared to the other months. The methods showed again that it is possible for the ANN models to forecast blooms higher than the magnitude they were trained with when predicting the 2015 bloom similar to the August 2015 prediction. The results for the two selection methods for October is shown below (Figure 18).

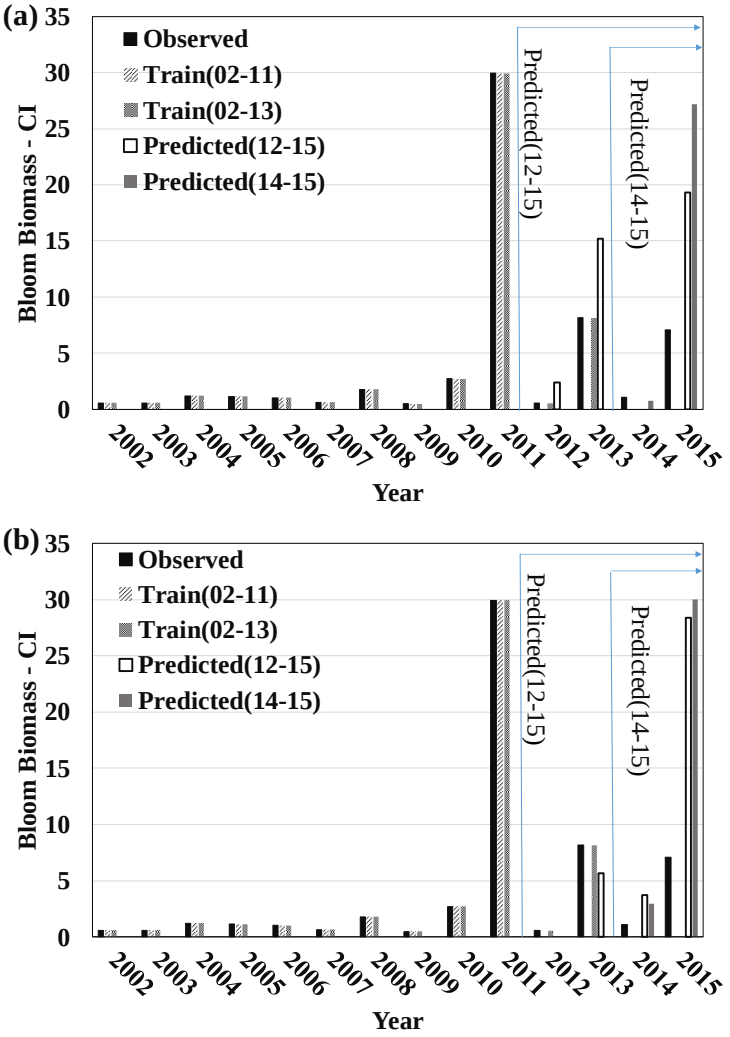


Figure 18. ANN Results for Both Selection Methods and Averaging Periods: (a) Spearman October and (b) Loading Period October

The predictions for October showed average results with the Spearman method over predicting two of the years and the loading period method only vastly overpredicting the 2015 bloom. Both the Spearman and loading period selection methods were unable to closely predict the 2015 bloom. The Spearman model overpredicted the bloom for 2013 however the loading period method was able to closely predict the bloom. The increased training had minor effects on both methods. The increased training for the loading period model had a small effect on the predictions however it followed the similar trend to the

Spearman model and caused the 2014 prediction to get slightly more accurate and the 2015 prediction to get less accurate.

The ANN models made for the two training periods and input selection methods were then plotted for their accuracy. The monthly trained and forecasted data for each of the four separate methods were combined and seperated for the training and forecasting data. The plots below (Figure 19) show the performance of ANN model trained to 2011 and 2013 using the loading period data selection method with correlation coefficients of 0.70 and 0.77, respectively.

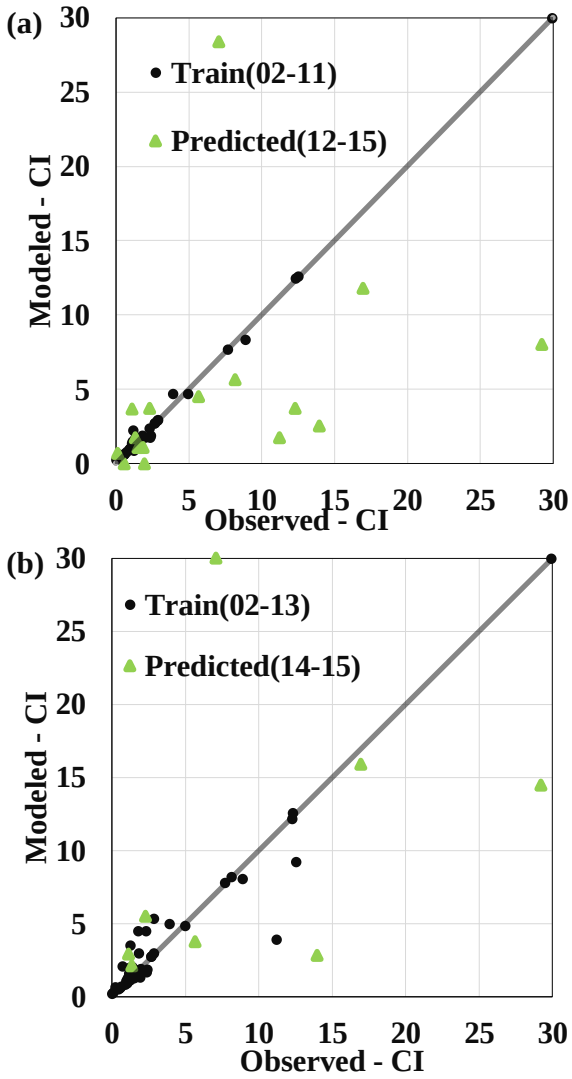


Figure 19. Performance of Two Loading Period Models: (a) Train(02-11) with 40 Trained and 16 Predicted and (b) Train(02-13) with 48 Trained and 8 Predicted

The loading period ANN model followed a similar trend to the Spearman model and underpredicted many of the higher magnitude blooms and overpredicted one. When updated to the longer training period, the loading period method and the Spearman method showed improvements in the overall accuracy for forecasting the larger blooms. The loading period, as well, only has two very underpredicted predictions yet it still had the overpredicted prediction. However, similar to the Spearman method for the extended training period, all of the predictions except one had the same bloom classification as the

observed booms. The plots below (Figure 20) show the accuracy of the Spearman selection methods trained to 2011 and 2013 with correlation coefficients of 0.79 and 0.83, respectively.

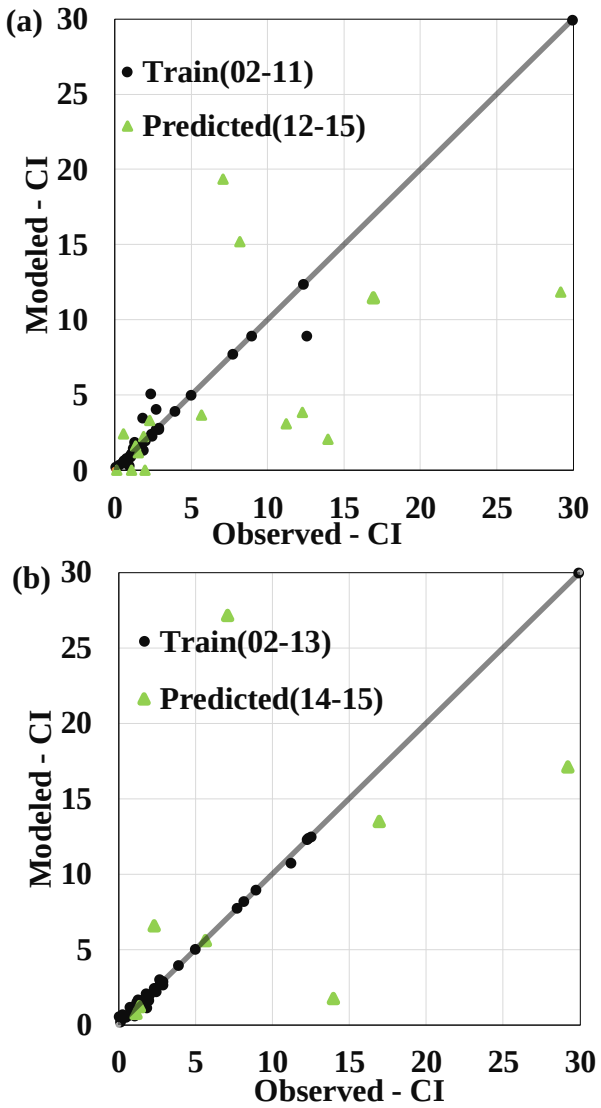


Figure 20. Performance of Two Spearman Models: (a) Train(02-11) with 40 Trained and 16 Predicted and (b) Train(02-13) with 48 Trained and 8 Predicted

The Spearman ANN model, when trained up to 2011, is able to forecast the lower magnitude blooms well. However, the model had all the predictions underpredicted

except two for the higher magnitude blooms. When increasing the training period to 2013, the new ANN model is also able to forecast the low magnitude blooms well and is able to forecast the higher magnitude blooms with improved accuracy when compared to the shorter training period. The model was unable to perfectly predict the high-class blooms however in many cases forecasted a similar class of bloom. In August, the 2015 bloom had a CI value of twenty-nine and all four prediction vastly underpredicted the bloom with the highest prediction being seventeen. However, when considering the class system, all four predictions and observed CI values for the 2015 bloom fall into the class 3 category. For the extended training period all of the predictions except one had the same bloom classification as the observed booms. Next, the variable importance was calculated for each of the months for both training periods and methods (Table 17, 18 and 19).

Table 17. July and August Spearman ANN Variable Importance for Both Training Periods.

July				**August**			
(02-11)		**(02-13)**		**(02-11)**		**(02-13)**	
PM6,1	18%	Q1,5	12%	SRP6,1	20%	Q4,2	24%
TKM1,6	17%	TKN1,6	10%	CI1,1	16%	PM4,2	20%
TP1,5	12%	Water3,4	9%	Q1,6	13%	SRP5,2	14%
Q1,6	11%	TP1,5	8%	Q4,2	12%	Q1,6	11%
PM1,5	6%	PM2,2	7%	Water1,2	10%	PM5,1	9%
PM2,2	5%	PM1,5	7%	SRP5,2	7%	CI1,1	9%
TKM3,3	5%	TKN5,2	7%	Water2,1	5%	Q5,1	6%
Q5,1	4%	TKN2,3	7%	Water2,2	5%	PM1,6	6%
Water3,4	5%	PM3,2	6%	Q5,1	4%	Water2,1	2%
PM3,1	4%	Q6,1	5%	PM1,6	3%		
TKM4,3	4%	Q2,2	5%	PM5,1	2%		
Q4,2	3%	Q3,1	4%	Water1,3	2%		
Q3,3	3%	Q4,1	4%	Water1,6	1%		
TKN5,2	3%	PM3,1	4%	TKN1,6	0%		
PM4,1	1%	Q3,2	3%				
		TP3,1	2%				

Table 18. September and October Spearman ANN Variable Importance for Both Training Periods.

September				**October**			
(02-11)		**(02-13)**		**(02-11)**		**(02-13)**	
Q6,1	17%	Q6,1	24%	Wind1,4	14%	CI1,1	28%
TKN1,1	13%	PM1,6	14%	Q1,6	13%	PM2,5	25%
TP3,4	11%	CI1,1	13%	TP1,6	12%	Wind4,1	20%
CI1,1	11%	PM3,1	11%	CI1,1	12%	Wind4,2	13%
PM3,1	8%	PM3,3	10%	PM5,2	11%	Q4,1	8%
PM1,6	8%	Q1,6	8%	TKN1,2	8%	TKN4,2	6%
TKN3,4	8%	Q3,1	7%	Air3,2	8%		
Q1,6	7%	TKN3,4	7%	Q5,2	7%		
PM3,3	7%	Q3,3	6%	PM1,6	4%		
Q3,1	6%			TKN4,2	4%		
Q3,3	6%			Air3,4	4%		
				Wind3,2	2%		

Table 19. Loading Period ANN Variable Importance for Both Training Periods.

	July		**August**		**September**		**October**	
Variable	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**	**(02-11)**	**(02-13)**
Q	43%	46%	33%	22%	31%	21%	27%	25%
TP	21%	26%	24%	5%	21%	25%	11%	7%
PM	11%	9%	21%	33%	32%	21%	24%	24%
SRP	25%	20%	13%	19%	3%	14%	8%	10%
PCI			9%	21%	14%	18%	30%	35%

NOTE: Empty boxes represent variables that were not considered.

The ANN variable importance for the Spearman method (Tables 17 and 18) showed interesting trends in the variable importance. For four of the eight Spearman models, Q showed the highest variable importance which is understandable as the flow of the Maumee River brings the nutrients into the lake. In most cases, the variables with a longer averaging period showed a higher importance. In the physical sense, this is understandable as a single lagged month is unable to shape the future blooms

singlehandedly outside of unusual cases such as in 2015. The previous month CI input variable is in the top four variable importance for all months it is considered except for the train(02-13) method in August where the CI input is showing as an overall valuable bloom prediction indicator.

The ANN variable importance for the loading period method (Table 19) also showed interesting results. Interestingly, Q is the only variable with a high importance in all eight methods with the lowest being 22%. PM also had a high importance for all methods outside of July with the lowest being 21%. These two variables, having the highest importance, are expected as they had high individual correlations for each of the bloom months.

CHAPTER V

SUMMARY AND CONCLUSION

5.1 Summary

HABs are a major problem all over the world particularly in Lake Erie. A widespread literature review was performed to understand the HAB problem in Lake Erie. The current methods to forecast HABs all over the world and specifically in Lake Erie were examined. An extensive literature review and analysis was performed on the possible variables for forecasting HABs. Two forecasting methods, CART and ANN, as well as two training periods and two input variable selection methods, nutrient loading period and Spearman's rank correlation were used. For the nutrient loading period selection method, only one set of input variables is used for forecasting whereas the Spearman selection method examines more variables than the nutrient loading period considering up to twenty-eight different averaging periods and lag times for each considered variable.

First, the CART models were tested with both classification methods, a 3-class and 5-class system resulting in the 3-class system being selected. The CART models were then created for both methods and training periods. Initially, when using the first training period of 2002 to 2011, the loading period method showed better precision in forecasting

HABs when compared to the Spearman selection method. When the training period was increased to 2013, both methods showed an improvement in the overall accuracy with Spearman having an 8.9% improvement and loading period a 5.4% improvement. However, with the extended training period, the loading period decision trees for August and September showed a slight increase in precision over the Spearman method and the Spearman method being slightly more precise in October.

After the CART models, the ANN models were created and analyzed for both selection methods and training periods. For both selection methods in the first training period the models often underpredicted the higher magnitude blooms. Many of the predictions did not predict the exact same CI as observed however in most cases both methods predicted the same class of bloom as observed. In most cases after increasing the training period to 2013, both ANN models improved their accuracy for predicting the higher magnitude of blooms. The correlation coefficient increased from 0.70 to 0.77 for the loading period selection method and from 0.79 to 0.83 for the Spearman selection method when extending the training period. Both input selection methods had some difficulty in predicting the 2015 HAB because the 2015 bloom was a special case in terms of nutrient loading as well as bloom time. There was a large amount of loading in June and July which is atypical. The 2015 bloom in July was 382% larger than any bloom recorded from 2002 to 2014. The monthly discharge for June was the highest recorded and third highest on record since the USGS started collecting data in 1939 (Stumpf, 2016). Similar to the CART method the ANN model showed an increase in accuracy when forecasting HABs with an extended training period.

5.2 Conclusions

Through the use of two machine learning techniques and selection methods, two forecasting models were created. For the nutrient loading period selection method, only one set of input variables is used for forecasting each month while the Spearman method uses different variables and periods for each month's forecast. The main advantage for the nutrient loading period selection method is that it allows for an ease of understanding which input variables and periods are used. The main advantage for the Spearman selection method is more accurate results however for the forecast to be completed for all four bloom months more data must be collected delaying the final forecasts.

The first machine learning technique, CART, has the main benefit of giving an advanced warning on the possible class of HABs months before they occur. When only considering nutrient contributing variables the CART model forecast for all four bloom months is completed by the end of June. The accuracy for the CART models to correctly classify the blooms increased greatly when extending the training period up to 2013. The CART models are valuable for watershed-planners and decision makers to prepare or change plans based on the class of blooms for each month. The CART models can make earlier forecasts for HABs when compared to the ANN models. One of the major limitations for the CART model however is only being able to forecast a class and not the specific peak CI value that is going to occur in each month.

The second machine learning technique, ANN, results are also valuable for watershed-planners and decision makers. The forecasts from the ANN models are able to predict the biomass of the future HABs. Forecasting the exact size of the bloom is valuable for any decision maker using the lake ranging from recreational use to

commercial fishing to water treatment plant managers. However, in a few cases, the ANN models vastly over predicted or underpredicted the HAB biomass. In most cases extending the training period for both input selection methods, the ANN models improved the accuracy of their predictions. When considering the class of blooms predicted the ANN models often predicted the correct class of bloom. One of the current limitations on the ANN models is the use of the previous month's CI value in the forecast. This limitation currently delays the forecast for the ANN models by up to three months. However, when ANN is used in conjunction with the CART model the CART model is able to make the early class forecast and allow the ANN model to give a more exact HAB biomass forecast closer to when the bloom occurs.

In future work, the ANN model can be improved to not include the previous month's CI value in order to make earlier forecasts. The extrapolation ability for both models can also be tested in order to forecast beyond the calibration range. The final ANN and CART models will be coded in a user interface system to forecast HABs in July, August, September, and October. This research was conducted to improve HAB forecasting while allowing consumers, recreational users, watershed planners, and decision makers to make more educated decisions and timely manage HABs in western Lake Erie. Lake Erie is an irreplaceable resource and this paper's aim is to help improve the accuracy of forecasting HABs as well as provide information to those affected by HABs occurring in Lake Erie.

REFERENCES

Anderson, C. R., Moore, S. K., Tomlinson, M. C., Silke, J., & Cusack, C. K. (2015). Living with harmful algal blooms in a changing world: strategies for modeling and mitigating their effects in coastal marine ecosystems. *Coastal and Marine Hazards, Risks, and Disasters. Elsevier BV, Amsterdam,* 495-561. doi:10.1016/b978-0-12-396483-0.00017-0.

Babovic, V., & Keijzer, M. (2000). Genetic programming as a model induction engine. *Journal of Hydroinformatics, 2*(1), 35-60.

Bernstein, M., Graham, R., Cline, D., Dolan, J. M., & Rajan, K. (2013, November). Learning-based event response for marine robotics. *In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 3362-3367. doi:10.1109/iros.2013.6696835.

Bertani, I., Obenour, D. R., Steger, C. E., Stow, C. A., Gronewold, A. D., & Scavia, D. (2016). Probabilistically assessing the role of nutrient loading in harmful algal bloom formation in western Lake Erie. *Journal of Great Lakes Research, 42*(6), 1184-1192. doi: 10.1016/j.jglr.2016.04.002.

Binding, C. E., Greenberg, T. A., & Bukata, R. P. (2013). The MERIS maximum chlorophyll index; its merits and limitations for inland water algal bloom monitoring. *Journal of Great Lakes Research, 39,* 100-107. doi: 10.1016/j.jglr.2013.04.005.

Blakey, T., Melesse, A. M., & Rousseaux, C. S. (2015). Toward connecting subtropical algal blooms to freshwater nutrient sources using a long-term, spatially distributed, in situ chlorophyll-a record. *Catena, 133,* 119-127. doi: 10.1016/j.catena.2015.05.001.

Bowden, G. J., Dandy, G. C., & Maier, H. R. (2005). Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology, 301*(1-4), 75-92. doi:10.1016/j.jhydrol.2004.06.021.

Bridgeman, T. B., Chaffin, J. D., & Filbrun, J. E. (2013). A novel method for tracking western Lake Erie Microcystis blooms, 2002–2011. *Journal of Great Lakes Research, 39*(1), 83-89. doi: 10.1016/j.jglr.2012.11.004.

Bullerjahn, G. S., McKay, R. M., Davis, T. W., Baker, D. B., Boyer, G. L., D'Anglada, L. V., & Wilhelm, S. W. (2016). Global solutions to regional problems: Collecting global expertise to address the problem of harmful cyanobacterial blooms. A Lake Erie case study. *Harmful Algae, 54,* 223-238. doi: 10.1016/j.hal.2016.01.003.

Burden, F., & Winkler, D. (2008). Bayesian Regularization of Neural Networks. *Methods in Molecular Biology™ Artificial Neural Networks, 458*, 23-42. doi:10.1007/978-1-60327-101-1_3.

Carmichael, W. W., & Board, H. P. A. (2013). *Human Health Effects from Harmful Algal Blooms: A Synthesis.* Ottawa: International Joint Commission.

Carvalho, G. A., Minnett, P. J., Banzon, V. F., Baringer, W., & Heil, C. A. (2011). Long-term evaluation of three satellite ocean color algorithms for identifying harmful algal blooms (Karenia brevis) along the west coast of Florida: A matchup assessment. *Remote sensing of environment, 115*(1), 1-18. doi: 10.1016/j.rse.2010.07.007.

Carvalho, G. A., Minnett, P. J., Fleming, L. E., Banzon, V. F., & Baringer, W. (2010). Satellite remote sensing of harmful algal blooms: A new multi-algorithm method for detecting the Florida Red Tide (Karenia brevis). *Harmful algae, 9*(5), 440-448. doi: 10.1016/j.hal.2010.02.002.

Cha, Y., Park, S. S., Kim, K., Byeon, M., & Stow, C. A. (2014). Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. *Water Resources Research, 50*(3), 2518-2532. doi:10.1002/2013wr014372.

Chattopadhyay, J., Sarkar, R. R., & Pal, S. (2004). Mathematical modelling of harmful algal blooms supported by experimental findings. *Ecological Complexity, 1*(3), 225-235. doi: 10.1016/j.ecocom.2004.04.001.

Cigizoglu, H. K. (2003). Estimation, forecasting and extrapolation of river flows by artificial neural networks. *Hydrological Sciences Journal, 48*(3), 349-361. doi:10.1623/hysj.48.3.349.45288.

Cortés, U., Sànchez-Marrè, M., Ceccaroni, L., R-Roda, I., & Poch, M. (2000). Artificial intelligence and environmental decision support systems. *Applied Intelligence, 13*(1), 77-91. doi: 10.1023/A:1008331413864.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178-3192. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Dolah, F. M. V., Roelke, D., & Greene, R. M. (2001). Health and ecological impacts of harmful algal blooms: risk assessment needs. *Human and Ecological Risk Assessment: An International Journal, 7*(5), 1329-1345. doi: 10.1080/20018091095032.

Eberhart, B. T. L., Bill, B. D., & Trainer, V. L. (2012). Remote sampling of harmful algal blooms: A case study on the Washington State coast. *Harmful algae, 19*, 39-45. doi: 10.1016/j.hal.2012.05.005.

Bingham M., Sinha S. K., & Lupi F. (2015). Economic Benefits of Reducing Harmful Algal Blooms in Lake Erie. *Environmental Consulting & Technology, Inc., Report*, pp 66.

Elshorbagy, A., Simonovic, S. P., & Panu, U. S. (2000). Performance evaluation of artificial neural networks for runoff prediction. *Journal of Hydrologic Engineering, 5*(4), 424-427. doi:10.1061/(asce)1084-0699(2000)5:4(424).

England, R., Ward, C. (2014). Lake Erie Protection Fund Final Reporting Sandusky Bay Algal Bloom Early Warning Study. *Erie County Health Department*.

Environmental Protection Agency. (2017). *Guidelines and Recommendations*. Retrieved February 5, 2017, from https://www.epa.gov/nutrient-policy-data/guidelines-and-recommendations.

Environmental Protection Agency. (2017). *Recommended Binational Phosphorus Targets.* Retrieved February 27, 2017, from https://www.epa.gov/glwqa/recommended-binational-phosphorus-targets.

Frassl, M. A., Rothhaupt, K. O., & Rinke, K. (2014). Algal internal nutrient stores feedback on vertical phosphorus distribution in large lakes. *Journal of Great Lakes Research, 40,* 162-172. doi: 10.1016/j.jglr.2013.11.001.

Frolov, S., Kudela, R. M., & Bellingham, J. G. (2013). Monitoring of harmful algal blooms in the era of diminishing resources: a case study of the US West Coast. *Harmful Algae, 21-22,* 1-12. doi: 10.1016/j.hal.2012.11.001.

Govindaraju, R. (2000). Artificial neural networks in hydrology: ii, hydrologic applications. *Journal of Hydrologic Engineering, 5*(2), 124-137. doi:10.1061/(asce)1084-0699(2000)5:2(124).

Govindaraju, R. S. (2000). Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115-123. doi:10.1061/(asce)1084-0699(2000)5:2(115).

Heidelberg, (2017). *National Center for Water Quality Research Tributary Data Download.* Retrieved March 01, 2017 from https://www.heidelberg.edu/academics/research-and-centers/national-center-for-water-quality-research/tributary-data-download.

Hettiarachchi, P., Hall, M. J., & Minns, A. W. (2005). The extrapolation of artificial neural networks for the modelling of rainfall—runoff relationships. *Journal of Hydroinformatics, 7*(4), 291-296.

Ho, J. C., & Michalak, A. M. (2015). Challenges in tracking harmful algal blooms: a synthesis of evidence from Lake Erie. *Journal of Great Lakes Research, 41*(2), 317-325. doi: 10.1016/j.jglr.2015.01.001.

Hudnell, H. K. (2010). The state of US freshwater harmful algal blooms assessments, policy and legislation. *Toxicon, 55*(5), 1024-1034. doi: 10.1016/j.toxicon.2009.07.021.

Indiana University–Purdue University Indianapolis. (2017). *What causes algal blooms?*. Retrieved February 15, 2017, from http://www.cees.iupui.edu/research/algal-toxicology/bloomfactors.

Jain, S. K., & Chalisgaonkar, D. (2000). Setting up stage-discharge relations using ANN. *Journal of Hydrologic Engineering, 5*(4), 428-433. doi:10.1061/(asce)1084-0699(2000)5:4(428).

Jian, L., Zhongwu, J., & Wenjun, Y. (2014). Numerical modeling of the Xiangxi River algal bloom and sediment-related process in China. *Ecological Informatics, 22*, 23-35. doi: 10.1016/j.ecoinf.2014.03.002.

Kang, H.Y., Rule, R.A., & Noble, P.A. (2011). Artificial Neural Network Modeling of Phytoplankton Blooms and its Application to Sampling Sites within the Same Estuary. *Treatise on Estuarine and Coastal Science,* 161-172. doi:10.1016/b978-0-12-374711-2.00908-6

Kasich, J., Butler, C., Zehringer, J., & Himes, L. (2012). State of Ohio Harmful Algal Bloom Response Strategy for Recreational Waters. *Department of Health, Environmental Protection Agency and Department of Natural Resources.*

Kim, K. S., & Park, J. H. (2009). A survey of applications of artificial intelligence algorithms in eco-environmental modelling. *Environmental Engineering Research, 14*(2), 102-110. doi:10.4491/eer.2009.14.2.102.

Kim, K., Park, M., Min, J. H., Ryu, I., Kang, M. R., & Park, L. J. (2014). Simulation of algal bloom dynamics in a river with the ensemble Kalman filter. *Journal of Hydrology, 519*, 2810-2821. doi: 10.1016/j.jhydrol.2014.09.073.

Kim, U., & Kaluarachchi, J. J. (2009). Hydrologic model calibration using discontinuous data: an example from the upper Blue Nile River Basin of Ethiopia. *Hydrological processes, 23*(26), 3705-3717. doi:10.1002/hyp.7465..

Kozacek, C. (2014, August 2). *Toledo issues emergency 'Do Not Drink Water' warning to residents*. Retrieved February 5, 2017, from http://www.circleofblue.org/2014/world/toledo-issues-emergency-warning-residents-drink-water/

Kozacek, C. (2014, April 9). *Cause of Lake Erie's harmful algal blooms gains more certainty*. Retrieved February 15, 2017, from http://www.circleofblue.org/2014/world/cause-lake-eries-harmful-algal-blooms-gains-certainty/

Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences, 5*, 89-97. doi:10.5194/adgeo-5-89-2005.

Kurekin, A. A., Miller, P. I., & Woerd, H. V. (2014). Satellite discrimination of Karenia mikimotoi and Phaeocystis harmful algal blooms in European coastal waters: Merged classification of ocean colour data. *Harmful Algae, 31*, 163-176. doi: 10.1016/j.hal.2013.11.003.

Lee, J. H., Huang, Y., Dickman, M., & Jayawardena, A. W. (2003). Neural network modelling of coastal algal blooms. *Ecological Modelling, 159*(2-3), 179-201. doi:10.1016/s0304-3800(02)00281-8.

Lee, J., Im, J., Kim, U., & Löffler, F. E. (2016). A data mining approach to predict in situ detoxification potential of chlorinated ethenes. *Environmental science & technology, 50*(10), 5181-5188. doi: 10.1021/acs.est.5b05090.

Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. *In Annual meeting of the society for academic emergency medicine in San Francisco, California* (pp. 1-14).

Lewitus, A. J., Horner, R. A., Caron, D. A., Garcia-Mendoza, E., Hickey, B. M., Hunter, M., ... & Lessard, E. J. (2012). Harmful algal blooms along the North American west coast region: History, trends, causes, and impacts. *Harmful Algae, 19*, 133-159. doi: 10.1016/j.hal.2012.06.009.

Li, H. M., Tang, H. J., Shi, X. Y., Zhang, C. S., & Wang, X. L. (2014). Increased nutrient loads from the Changjiang (Yangtze) River have led to increased Harmful Algal Blooms. *Harmful Algae, 39*, 92-101. doi: 10.1016/j.hal.2014.07.002.

Limoges, A., de Vernal, A., & Ruiz-Fernández, A. C. (2015). Investigating the impact of land use and the potential for harmful algal blooms in a tropical lagoon of the Gulf of Mexico. *Estuarine, Coastal and Shelf Science, 167*, 549-559. doi: 10.1016/j.ecss.2015.11.005.

Lou, I., Xie, Z., Ung, W. K., & Mok, K. M. (2015). Integrating Support Vector Regression with Particle Swarm Optimization for numerical modeling for algal blooms of freshwater. *Applied Mathematical Modelling, 39*(19), 5907-5916. doi: 10.1016/j.apm.2015.04.001.

Lou, X., & Hu, C. (2014). Diurnal changes of a harmful algal bloom in the East China Sea: Observations from GOCI. *Remote Sensing of Environment, 140*, 562-572. doi: 10.1016/j.rse.2013.09.031.

Lourakis, M. I. (2005). A brief description of the Levenberg-Marquardt algorithm implemented by levmar. *Foundation of Research and Technology, 4*(1).

Lui, G. C., Li, W. K., Leung, K. M., Lee, J. H., & Jayawardena, A. W. (2007). Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter. *Ecological modelling, 200(1-2),* 130-138. doi: 10.1016/j.ecolmodel.2006.06.017.

MathWorks. (2017). *Fitctree*. Retrieved March 01, 2017 from https://www.mathworks.com/help/stats/fitctree.html#References.

MathWorks. (2017). *Trainbr*. Retrieved March 01, 2017 from https://www.mathworks.com/help/nnet/ref/trainbr.html?s_tid=srchtitle

Mao, J., Jiang, D., & Dai, H. (2015). Spatial–temporal hydrodynamic and algal bloom modelling analysis of a reservoir tributary embayment. *Journal of Hydro-environment Research, 9*(2), 200-215. doi: 10.1016/j.jher.2014.09.005.

Messiana, I. (2016, June 6). *Official links water costs to past mayors*. Retrieved February 13, 2017, from http://www.toledoblade.com/Politics/2016/06/06/Official-links-water-costs-to-past-mayors.html.

Messiana, I. (2017, January 2). *Water rates rise 13.2% in '17 for 4th consecutive year.* Retrieved February 13, 2017, from http://www.toledoblade.com/local/2017/01/02/Toledo-Water-rates-rise-13-2-percent-in-2017-for-4th-consecutive-year.html.

Millie, D. F., Weckman, G. R., Fahnenstiel, G. L., Carrick, H. J., Ardjmand, E., Young, W. A., ... & Shuchman, R. A. (2014). Using artificial intelligence for CyanoHAB niche modeling: discovery and visualization of Microcystis–environmental associations within western Lake Erie. *Canadian Journal of Fisheries and Aquatic Sciences, 71*(11), 1642-1654. doi:10.1139/cjfas-2013-0654.

Misra, A. K., Chandra, P., & Raghavendra, V. (2011). Modeling the depletion of dissolved oxygen in a lake due to algal bloom: Effect of time delay. *Advances in Water Resources, 34*(10), 1232-1238. doi: 10.1016/j.advwatres.2011.05.010.

Mitra, A., & Flynn, K. J. (2010). Modelling mixotrophy in harmful algal blooms: More or less the sum of the parts? *Journal of Marine Systems, 83*(3), 158-169. doi: 10.1016/j.jmarsys.2010.04.006.

Muttil, N., & Chau, K. W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution, 28*(3-4), 223-238. doi:10.1504/ijep.2006.011208.

Neilson, M., L'Italien, S., Glumac, V., Williams, D., & Bertram, P. (1995). state of the lakes ecosystem conference background paper. Nutrients: Trends and system response. *United States Environmental Protection Agency EPA 905-R-95-015*.

Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., ... & Zechman, E. (2010). State of the art for genetic algorithms and beyond in water resources planning and management. *Journal of Water Resources Planning and Management, 136*(4), 412-432. doi:10.1061/(asce)wr.1943-5452.0000053.

Norton, L., Elliott, J. A., Maberly, S. C., & May, L. (2012). Using models to bridge the gap between land use and algal blooms: An example from the Loweswater catchment, UK. *Environmental Modelling & Software, 36*, 64-75. doi: 10.1016/j.envsoft.2011.07.011.

Obenour, D. R., A. D. Gronewold, C. A. Stow, & D. Scavia (2014). Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. *Water Resources Research, 50*(10), 7847–7860. doi:10.1002/2014wr015616.

Oliver, A. A., Dahlgren, R. A., & Deas, M. L. (2014). The upside-down river: Reservoirs, algal blooms, and tributaries affect temporal and spatial patterns in nitrogen and phosphorus in the Klamath River, USA. *Journal of Hydrology, 519*, 164-176. doi: 10.1016/j.jhydrol.2014.06.025.

Ohio Department of Health. (2016). *Harmful Algal Blooms*. Retrieved March 01, 2017, from https://www.odh.ohio.gov/odhprograms/eh/HABs/algalblooms.aspx.

Painsky, A., & Rosset, S. (2016). Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1. doi:10.1109/tpami.2016.2636831.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote sensing of environment, 86*(4), 554-565. doi:10.1016/s0034-4257(03)00132-9.

Park, Y., Cho, K. H., Park, J., Cha, S. M., & Kim, J. H. (2015). Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of the Total Environment, 502*, 31-41. doi: 10.1016/j.scitotenv.2014.09.005.

Park, Y., Pachepsky, Y. A., Cho, K. H., Jeon, D. J., & Kim, J. H. (2015). Stressor–response modeling using the 2D water quality model and regression trees to predict chlorophyll-a in a reservoir system. *Journal of Hydrology, 529*, 805-815. doi: 10.1016/j.jhydrol.2015.09.002.

Perri, K. A., Sullivan, J. M., & Boyer, G. L. (2015). Harmful algal blooms in Sodus Bay, Lake Ontario: A comparison of nutrients, marina presence, and cyanobacterial toxins. *Journal of Great Lakes Research, 41*(2), 326-337. doi: 10.1016/j.jglr.2015.03.022.

Raine, R., McDermott, G., Silke, J., Lyons, K., Nolan, G., & Cusack, C. (2010). A simple short range model for the prediction of harmful algal events in the bays of southwestern Ireland. *Journal of Marine Systems, 83*(3), 150-157. doi: 10.1016/j.jmarsys.2010.05.001.

Rast, W., Lee, G. F., Corvallis Environmental Research Laboratory., & Organisation for Economic Co-operation and Development. (1978). *Summary analysis of the North American (US portion) OECD eutrophication project: Nutrient loading--lake response relationships and trophic state indices.* Corvallis, Ore: Corvallis Environmental Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency.

Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree

(CART) models. *Expert Systems with Applications, 29*(1), 65-74. doi: 10.1016/j.eswa.2005.01.006.

Recknagel, F., French, M., Harkonen, P., & Yabunaka, K. I. (1997). Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling, 96*(1-3), 11-28. doi:10.1016/s0304-3800(96)00049-x.

Reutter, J., & Dierkes, C. (2014, August 5). *Harmful Algal Bloom Q&A and Updates*. Retrieved March 01, 2017, from https://ohioseagrant.osu.edu/news/2014/a8990/harmful-algal-bloom-qa-updates

Shukla, J. B., Misra, A. K., & Chandra, P. (2008). Modeling and analysis of the algal bloom in a lake caused by discharge of nutrients. *Applied Mathematics and Computation, 196*(2), 782-790. doi: 10.1016/j.amc.2007.07.010.

Simon, A., & Shanmugam, P. (2012). An algorithm for classification of algal blooms using MODIS-Aqua data in oceanic waters around India. *Advances in Remote Sensing, 01*(02), 35-51. doi:10.4236/ars.2012.12004.

Singh, S., Vashishtha, V., & Singla, T. (2014). Artificial Neural Network. *International Journal of Research, 1*(9), 934-942.

Sivapragasam, C., Muttil, N., Muthukumar, S., & Arun, V. M. (2010). Prediction of algal blooms using genetic programming. *Marine pollution bulletin, 60*(10), 1849-1855. doi: 10.1016/j.marpolbul.2010.05.020.

Yao, J., Xiao, P., Zhang, Y., Zhan, M., & Cheng, J. (2011). A mathematical model of algal blooms based on the characteristics of complex networks theory. *Ecological modelling, 222*(20-22), 3727-3733. doi: 10.1016/j.ecolmodel.2011.09.006.

Smayda, T. J. (2008). Complexity in the eutrophication–harmful algal bloom relationship, with comment on the importance of grazing. *Harmful Algae, 8*(1), 140-151. doi: 10.1016/j.hal.2008.08.018.

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. doi:10.2489/jswc.70.2.27a.

Solé, J., Estrada, M., & Garcia-Ladona, E. (2006). Biological control of harmful algal blooms: A modelling study. *Journal of Marine Systems, 61*(3-4), 165-179. doi: 10.1016/j.jmarsys.2005.06.004.

Song, W., Dolan, J. M., Cline, D., & Xiong, G. (2015). Learning-Based Algal Bloom Event Recognition for Oceanographic Decision Support System Using Remote Sensing Data. *Remote Sensing, 7*(10), 13564-13585. doi:10.3390/rs71013564.

Srivastava, A., Singh, S., Ahn, C. Y., Oh, H. M., & Asthana, R. K. (2013). Monitoring approaches for a toxic cyanobacterial bloom. *Environmental science & technology, 47*(16), 8999-9013. doi:10.1021/es401245k.

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., ... Burtner, A. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae, 54*, 160-173. doi: 10.1016/j.hal.2016.01.005.

Stumpf, R. P., Johnson, L. T., Wynne, T. T., & Baker, D. B. (2016). Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie. *Journal of Great Lakes Research, 42*(6) 1174-1183. doi: 10.1016/j.jglr.2016.08.006.

Stumpf, R. P., Wynne, T. T., Baker, D. B., & Fahnenstiel, G. L. (2012). Interannual variability of cyanobacterial blooms in Lake Erie. *PLoS One, 7*(8). doi: 10.1371/journal.pone.0042444.

Sunda, W. G., & Shertzer, K. W. (2014). Positive feedbacks between bottom-up and top-down controls promote the formation and toxicity of ecosystem disruptive algal blooms: A modeling study. *Harmful Algae, 39*, 342-356. doi: 10.1016/j.hal.2014.09.005.

Thirumalaiah, K., & Deo, M. C. (2000). Hydrological forecasting using neural networks. *Journal of Hydrologic Engineering, 5*(2), 180-189. doi: 10.1061/(ASCE)1084-0699(2000)5:2(180).

Tokar, A. S., & Markus, M. (2000). Precipitation-runoff modeling using artificial neural networks and conceptual models. *Journal of Hydrologic Engineering, 5*(2), 156-161. doi:10.1061/(asce)1084-0699(2000)5:2(156).

Velo-Suárez, L., & Gutiérrez-Estrada, J. C. (2007). Artificial neural network approaches to one-step weekly prediction of Dinophysis acuminata blooms in Huelva (Western Andalucía, Spain). *Harmful Algae, 6*(3), 361-371. doi: 10.1016/j.hal.2006.11.002.

Villacorte, L. O., Tabatabai, S. A. A., Anderson, D. M., Amy, G. L., Schippers, J. C., & Kennedy, M. D. (2015). Seawater reverse osmosis desalination and (harmful) algal blooms. *Desalination, 360*, 61-80. doi: 10.1016/j.desal.2015.01.007.

Walsh, J. J., Penta, B., Dieterle, D. A., & Bissett, W. P. (2001). Predictive ecological modeling of harmful algal blooms. *Human and Ecological Risk Assessment: An International Journal, 7*(5), 1369-1383. doi:10.1080/20018091095069.

Wells, M. L., Trainer, V. L., Smayda, T. J., Karlson, B. S., Trick, C. G., Kudela, R. M., ... & Cochlan, W. P. (2015). Harmful algal blooms and climate change: Learning from the past and present to forecast the future. *Harmful algae, 49*, 68-93. doi: 10.1016/j.hal.2015.07.009.

Wynne, T. T., & Stumpf, R. P. (2015). Spatial and temporal patterns in the seasonal distribution of toxic cyanobacteria in western Lake Erie from 2002–2014. *Toxins, 7*(5), 1649-1663. doi:10.3390/toxins7051649.

Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., & Dyble, J. (2010). Characterizing a cyanobacterial bloom in western Lake Erie using satellite imagery and meteorological data. *Limnology and Oceanography, 55*(5), 2025-2036. doi:10.4319/lo.2010.55.5.2025.

Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., Warner, R. A., Tester, P. A., Dyble, J., & Fahnenstiel, G. L. (2008). Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes. *International Journal of Remote Sensing, 29*(12), 3665-3672. doi:10.1080/01431160802007640.

Zeng, W., Song, Q., Liu, H., & Wang, T. (2010). Research on ANN-based pre-warning water bloom model of LiuHai Lake in Beijing. *Procedia Environmental Sciences, 2*, 625-635. doi: 10.1016/j.proenv.2010.10.070.

Zhao, J., Temimi, M., & Ghedira, H. (2015). Characterization of harmful algal blooms (HABs) in the Arabian Gulf and the Sea of Oman using MERIS fluorescence data. *ISPRS Journal of Photogrammetry and Remote Sensing, 101*, 125-136. doi: 10.1016/j.isprsjprs.2014.12.010.