5-2021

A Review of Harmful Algal Bloom Prediction Models for Lakes and Reservoirs

Jade Snyder Echard
Utah State University

A REVIEW OF HARMFUL ALGAL BLOOM PREDICTION MODELS FOR LAKES AND RESERVOIRS

by

Jade Snyder Echard

A project report submitted in partial fulfillment
of the requirements for the degree
of

MASTERS OF SCIENCE

in

Environmental Engineering

Approved:

David K. Stevens, PhD Major Professor	Joan E. McLean, MS Committee Member

R. Ryan Dupont, PhD
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah
2020

Abstract

Anthropogenic activity has led to eutrophication in water bodies across the world. This eutrophication promotes blooms, cyanobacteria being among the most notorious bloom organisms. Cyanobacterial blooms (more commonly referred to as harmful algal blooms (HABs)) can devastate an ecosystem. Cyanobacteria are resilient microorganisms that have adapted to survive under a variety of conditions, often outcompeting other phytoplankton. Some species of cyanobacteria produce toxins that ward off predators. These toxins can negatively affect the health of the aquatic life, but also can impact animals and humans that drink or come in contact with these noxious waters. Although cyanotoxin's effects on humans are not as well researched as the growth, behavior, and ecological niche of cyanobacteria, their health impacts are of large concern. It is important that research to mitigate and understand cyanobacterial blooms and cyanotoxin production continues. This project supports continued research by addressing an approach to collect and summarize published articles that focus on techniques and models to predict cyanobacterial blooms with the goal of understanding what research has been done to promote future work. The following report summarizes 34 articles from 2003 to 2020 that each describe a mechanistic or data driven model developed to predict the occurrence of cyanobacterial blooms or the presence of cyanotoxins in lakes or reservoirs with similar climates to Utah. These articles showed a shift from more mechanistic approaches to more data driven approaches with time. This resulted in a more individualistic approach to modeling, meaning that models are often produced for a single lake or reservoir and are not easily comparable to other models for different systems.

Acknowledgements

I would like to thank my major professor, Dr. David Stevens for his help and guidance on this project. I would like to thank my committee members, Professor Joan McLean and Dr. Ryan Dupont for their help and support. I would additionally like to thank the Utah Water Research Lab (UWRL) for the financial support to proceed with this project.

Table of Contents

Table of Tables

Table of Figures

Introduction

Lakes and reservoirs in Utah are negatively impacted by algal blooms every summer. Utah Lake is the most notorious for its large blooms that impact the community's ability to use its water for recreation or irrigation. Algal blooms produce many odor, taste, and aesthetic problems in lakes with both recreational and drinking water beneficial uses. Reservoirs used for drinking water are of special concern because many cyanobacteria, one of the main microorganism groups responsible for blooms, produce toxins. These toxins can cause a host of different harmful health effects in humans and animals. Water treatment facilities may need to change their operational procedures with noxious waters to ensure toxins do not enter the distribution system. Whether a bloom produces toxins or not, it can upset an ecosystem, negatively impacting life forms from other phytoplankton to humans. It is important that progress is made to understand and mitigate the effects of these harmful algal blooms.

Many mathematical models have been produced that use environmental parameter inputs to understand which factors influence blooms and to predict when blooms will occur. The purpose of this report is to aggregate recent literature on bloom prediction methods. This review will be beneficial to further research as it will stand as a baseline for what has been done in the past so that research is progressed and not repeated. Advancing research in this field will help facilitate predicting and mitigating the formation of algal blooms.

Background

Cyanobacteria are the most notorious microorganisms that contribute to harmful algal blooms in bodies of fresh water (CDC 2020). This section describes cyanobacteria and discusses their effects on the environment, the toxins they produce with related health effects, and their growth rates.

Cyanobacteria

Cyanobacteria (historically called blue-green algae) are prokaryotic phytoplankton that form the base of the aquatic food chain alongside algae, diatoms, and dinoflagellates. Cyanobacteria were originally classified as algae due to their aquatic presence and ability to utilize light, carbon dioxide (CO_2), and other inorganic and organic nutrients to photosynthetically produce oxygen (Paerl et al. 2001). Most literature uses both the terms cyanobacteria and blue-green algae synonymously. It was not until the 1950s and the invention of the electron microscope that clear differences between eukaryotic and prokaryotic cells were found and cyanobacteria were properly classified as bacteria (Brock 1979).

Although recently reclassified, these ancient phototrophic microorganisms have been producing oxygen for about 2.5 billion years (Paerl et al. 2001). Cyanobacteria have evolved and adapted to thrive in various conditions of light, temperature, depth, and nutrient levels. According to the Redfield ratio (Redfield 1934), a balanced aquatic system has an N:P atomic ratio of 16:1. If a given body of water has a ratio less than 16 then nitrogen is limiting growth and if it is larger than 16 phosphorus is limiting growth (Rhee 1982). However, Rhee (1982) notes that the N:P ratio is a general value created for oceanic environments as a whole, and variation in species and environment may operate optimally under different ratios. In conditions where nitrogen is limited, most species of cyanobacteria can fix nitrogen (N_2) from the atmosphere. In conditions where phosphorous input is intermittent or limited, they can take up excess phosphorus and store it. Thus, cyanobacteria can flourish in both depleted or excessive levels of nutrients. While some species of cyanobacteria float on the surface of a water body and some reside inches to feet below the surface, some species of cyanobacteria have gas vacuoles to regulate buoyancy in order to absorb the optimum amount of light for growth. Adjusting buoyancy allows cyanobacteria to survive when waters become turbid and also allows nutrients to be used from different layers of the water column. Many species of cyanobacteria produce toxins that negatively affect predators and competitors, allowing them to grow without predatory or competitive control. Cyanobacteria can also survive in low levels of CO_2 and high pH conditions because they can synthesize carbonic anhydrase, an enzyme that can produce CO_2 from bicarbonate (HCO_3^-) (Paerl 1982). All of these adaptations allow cyanobacteria to dominate other phytoplankton in many situations.

Although they are able to survive in various conditions, cyanobacteria thrive with rapid growth rates in conditions with low grazing rates, little to no mixing, long residence times, and high nutrient loading. Human-influenced eutrophication and climate change have aided in providing nutrient rich environments for cyanobacteria to exploit. In the United States, 21% of lakes and reservoirs are considered to be hypereutrophic (EPA 2017). Rapidly reproducing cyanobacteria lead to cyanobacterial blooms, one type of the phenomena that have come to be commonly referred to as harmful algal blooms (HABs), which are a global occurrence, negatively affecting waterbodies from rivers to reservoirs to oceans. HABs can result from over production of other phytoplankton, but blooms caused by cyanobacteria are the most notorious and problematic.

When HABs occur, they negatively affect the ecosystem and have adverse effects on any beneficial use of the water source. HABs create unsightly aesthetic, odor, taste, and toxicity problems. Blooms can deplete the water body of important nutrients other organisms need, causing a sudden increase in cyanobacterial biomass and a simultaneous sharp decline in all other forms of biomass due to the toxins that inhibit their growth. The decay of this biomass results in depletion of oxygen in the underlying water leading to biological and chemical changes like hypoxia, anoxia, toxic hydrogen sulfide release, mobilization of some heavy metals, and even release of sediment nutrients facilitating further eutrophication (Paerl et al. 2001). When noxious species of cyanobacteria are present in blooms, they can release toxins harmful to other phytoplankton, zooplankton, fish, animals, and even humans if the source is being used for recreation or drinking water. A well-known example of the severe problems HABs can cause is when Toledo, Ohio had to issue a "do not drink" advisory to more than 400,000 residents for Lake Erie's water, plagued annually with blooms (Steffen 2017). Cyanobacterial toxins from Lake Erie had overwhelmed the city's water treatment facility and concentrations greater than the World Health Organization's (WHO) drinking water guidelines of 1 µg/L of microcystins were found in their finished water (Steffen 2017). This toxin production is the main cause of concern for human use of water containing cyanobacteria.

Cyanotoxins

Toxicity from cyanobacteria comes in many different forms and from many different species. About 40 species have been found to be toxic (Carmichael 2001). The most prominent and well researched genera include *Aphanizomenon*, *Dolichospermum* (*Anabaena*), *Cylindrospermopsis*, *Gloeotrichia*, *Nodularia*, *Microcystis*, *Oscillatoria* (*Planktothrix*), and *Lyngbya*. The three main genera found in Utah fresh waters are *Aphanizomenon*, *Dolichospermum* (Anabaena), and *Microcystis* (UDEQ 2019). Each cyanobacterial taxon has different qualities in terms of toxicity, N_2 fixation, and buoyancy (Table 1). Toxins produced by cyanobacteria or cyanotoxins include neurotoxins, hepatotoxins, and dermatoxins.

Table 1. Cyanobacterial Qualities and Toxin Production (Paerl et al. 2001, Carmichael 2001, EPA 2020, UDEQ 2020)

	Able to fix N_2?	Have buoyancy control?	Neurotoxins		Hepatotoxins			Dermatoxins
			Anatoxins	Saxitoxins	Cylindrospermopsins	Microcystins	Nodularins	Lyngbyatoxin
Aphanizomenon	Yes	Yes	X	X	X	X*		
Dolichospermum (Anabaena)	Yes	Yes	X	X	X**	X		
Cylindrospermospsis	Yes	No	X**	X*	X			
Gleotrichia	Yes	Yes				X*		
Nodularia	Yes	Yes				X*	X*	
Microcystis	No	Yes	X*			X		
Oscillatoria (Planktothrix)	No	Yes	X	X**		X		
Lygbya	No	No		X	X**			X*

*Only confirmed by one of three sources
**Only confirmed by two of three sources

Neurotoxins affect the nervous system, hepatotoxins damage the liver, and dermatoxins damage skin and mucous membranes. These toxins affect other phytoplankton, zooplankton, and fish that live in the same water, but can additionally bioaccumulate in the aquatic food chain. Cyanotoxins have been known to cause illness or death in wild and domesticated animals that bathed in, consumed, or breathed in the toxins (Carmichael 2001). In humans, testing and research is more limited, but many health-related symptoms have been attributed to cyanotoxins. Certain cyanotoxins are believed to cause liver cancer and promote tumors, and there have been confirmed deaths in Brazil where cyanotoxins were introduced through medical dialysis (Carmichael 2001).

Little is definitive about the effect of cyanotoxins on humans, but because the risks associated with exposure are so high, regulations are being set across the world to protect the public. In Utah, the Department of Environmental Quality (DEQ) (2020) has put a monitoring system in place to issue warning and danger advisories to the public. Three toxins, common in Utah lakes and reservoirs, are monitored at a warning and danger level as shown in Table 2.

Table 2. Cyanotoxin Exposure Warning and Danger Levels (UDEQ 2020)

<table>
<tr><th>Toxin</th><th>Warning Advisory</th><th>Danger Advisory</th></tr>
<tr><td>Microcystins (µg/L)</td><td>8</td><td>2,000</td></tr>
<tr><td>Cylindrospermopsin (µg/L)</td><td colspan="2">15</td></tr>
<tr><td>Anatoxin-a (µg/L)</td><td>15</td><td>90</td></tr>
</table>

The DEQ notes that the public could be exposed to cyanotoxins by ingestion, skin contact, and inhalation. At the warning advisory level, DEQ cautions that there is potential for long-term illness as well as short-term effects such as skin and eye irritation, nausea, vomiting, and diarrhea. At the danger advisory level, they caution that there is potential for acute poisoning, long-term illness, and the same short-term effects (UDEQ 2020). Data for differentiation breakpoints between a warning and danger advisory is less available for cylindrospermopsin, therefore, 15 mg/L has been set as the overall cyanotoxin advisory threshold.

Utah has not yet established maximum contaminant levels for cyanotoxins in drinking water, so the state encourages municipalities to remain within compliance of the Health Advisory Levels (HALs) established by the EPA (2020). At the levels shown in Table 3, utilities are encouraged to increase monitoring, change treatment strategies, and issue a notification of "do not drink/do not boil" advisories (UDEQ 2018). The EPA created this chart for two population groups, children under 6 years old and adults/children 6 years and older, establishing that no adverse effect to humans should happen at these levels for up to 10 days. Building from these suggested safety limits, the DEQ additionally suggests that these limits should be applied to more susceptible adults and other adults as Table 3 shows.

Table 3. Cyanotoxin Health Advisories (EPA 2020, UDEQ 2020)

10-Day Health Advisories	**Levels (µg/L)**
Microcystins	
Children under 6 years old as well as pregnant women, nursing mothers, elderly, immune-compromised, and dialysis patients	0.3
Children 6 years and older and other adults	1.6
Cylindrospermopsin	
Children under 6 years old as well as pregnant women, nursing mothers, elderly, immune-compromised, and dialysis patients	0.7
Children 6 years and older and other adults	3.0

Predicting HABs

The health effects associated with being exposed to cyanotoxins are being taken seriously around the world. It is becoming increasingly important to develop better models and monitoring systems to predict when HABs are likely to occur in order to mitigate the blooms and better warn utilities and the public. There have been modeling approaches that use environmental conditions to predict cyanotoxin concentrations (Alonso Fernández et al. 2013, García Nieto et al. 2011), but most modeling approaches are based on or are used to predict cyanobacterial growth rates and concentrations.

Mechanistic Models versus Data-Driven Models

Existing models used to predict HABs fall into two main categories: mechanistic or data-driven. Guven and Howard (2006) reviewed cyanobacteria models for both lake and river systems published between 1974 and 2003. The review covered mechanistic models for growth and movement of cyanobacteria in lakes and rivers and data-driven models (specifically artificial neural networks) for rivers. The current review builds on the 2006 review and focuses on cyanobacteria prediction models produced for lake or reservoir systems.

Mechanistic Models

Mechanistic models are based on fundamental laws of physical, biological, and chemical processes in nature, and use sets of algebraic and differential equations to describe these processes. In cyanobacteria prediction models, kinetic growth equations have been developed to understand how cyanobacteria grow quickly and take over bodies of water. Many of the kinetic growth equations use input parameters such as pH, light intensity, ionic conditions, temperature, nutrients, and other environmental factors, and produce a relationship to a specific growth rate. These kinetic equations can be used to understand growth of cyanobacteria but cannot necessarily be used to understand how harmful or toxic a bloom will be.

Growth rate equations have often been created by simply multiplying functions of limiting factors that growth is dependent upon:

$$\mu = f(N) * f(I) * f(T) \tag{1}$$

where μ (time^{-1}) is the cyanobacterial specific growth rate and $f(N),\ f(I),\ f(T)$ are functions representing the effects of nutrients, irradiance, and temperature on the growth rate (Giannuzzi 2019).

Although growth is dependent on several factors, often (and increasingly) growth equations are formulated based on a single limiting or dependent parameter. Liebig's Law of Minimum posits that growth is not controlled by each parameter needed for growth but is controlled by the one parameter that is deficient relative to all other requirements (ACG 2020), as illustrated in Figure 1. If the other growth factors are in abundance, growth is often regulated by the limiting growth parameter.

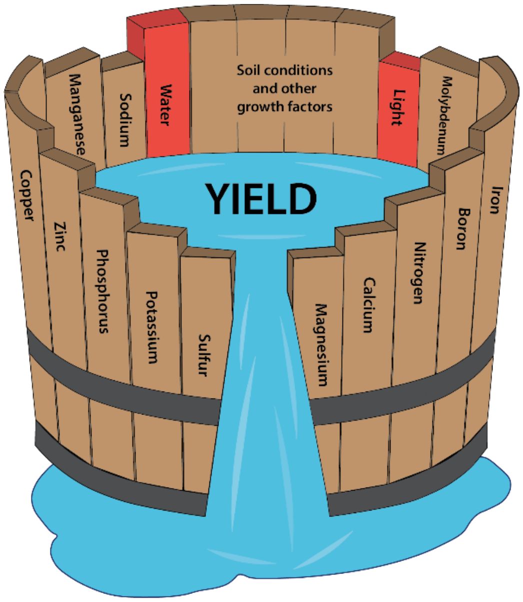

Figure.1 Illustration of Liebig's Law of the Minimum (Adapted from ACG 2020)

Rhee (1982) argues that limitation is not often as simple as Liebig's principle. He states that multiple simultaneous limitations are a more accurate representation because systems, inputs, and conditions are constantly changing. Guven and Howard (2006) outline several growth rate equations that use both Giannuzzi's equation and the idea of a limiting growth factor to create kinetic equations more specific to certain parameters or conditions. The studies they summarize focus on growth rates in areas where there is high or low irradiance, nutrient saturation or depletion, and turbidity or very still waters. In different conditions, kinetic growth equations become more specific using parameters such as buoyancy, density, velocity, respiration rate, birth rate, and death rate to produce accurate growth rates (Guven and Howard 2006).

Data Driven Models

Data-driven models determine relationships based on input data rather than on theoretical principles. These models predict cyanobacteria through specified empirical relationships or by machine learning techniques. Data-driven models require larger amounts of data to be collected and are often designed for a specific system and, unlike mechanistic models, are not likely to be universally applicable. Specific data-driven models are discussed in the results sections. Guven and Howard's (2006) review only looked at the machine learning approach known as artificial neural networks in river systems, while this review focuses on data-driven prediction models specifically in lake and reservoir systems.

Mechanistic modeling and understanding growth rates support a better understanding of cyanobacteria to forecast when HABs are likely to occur. Data-driven modeling uses data from a specific system to understand parameters that lead to HABs in the past to predict HABs in the future, with no intention of understanding the underlying science. Research on predicting and monitoring HABs is plentiful and has been done for decades. However, it is important that research is continually advanced as blooms are still a large environmental problem today.

Objectives

The overall objective of this project was to collect qualitative ideas, methods, and models about cyanobacterial blooms and cyanotoxin predictions from various recent (> 2003) studies and combine that work into a summarized report to facilitate further studies, experiments, and models. This report provides a consolidated review of cyanobacteria models published since the review by Guven and Howard (2006) (~2003 to present) to facilitate research that can provide accurate predictions of cyanobacteria and the production of cyanotoxins to help reduce the human and environmental health risks.

This objective was met by achieving smaller sequential goals including: compiling key words from selected previous studies, using those key words to search for related studies, organizing the studies based on relevance to the project, reading each relevant study to accumulate valuable information, and combining the knowledge from those studies into one report. A key emphasis for this report was to focus on mathematical models produced for lake and reservoir systems with ecological and environmental conditions similar to Utah.

This study is important to further research and development of cyanobacteria mitigation strategies and tactics and ultimately minimize the growth of cyanobacteria in Utah lakes. Utah lakes are highly susceptible to cyanobacterial blooms due to the state's phosphorous rich geology, agricultural practices, temperate climate, weather conditions, and plentiful slow-moving reservoirs (UDEQ 2020, Randall et al. 2019). More needs to be done to protect Utah's waterbodies.

Materials and Methods

To create a comprehensive review, referenced journal articles related to the topic of modeling and predicting cyanobacterial blooms and toxin production were collected from the 1950s to the most recent studies published in 2020. Papers were collected using scientific databases of peer reviewed articles. Sixty-four papers published prior to 2016 had been previously identified that consider cyanobacterial growth and modeling, and these papers were used to gather a set of relevant keywords so that a thorough search of the database could be performed. The keywords taken from each of the 64 papers were organized and categorized into low, moderate, or high relevance; for example, specific lake names were given a low rating, specific model names were given a moderate rating, and a term such as "prediction models" was given a high rating.

The keywords were then combined and input into the widely used scientific literature database search engine, Scopus (Elsevier 2020), to locate articles that support this review. Each search included different combinations or synonyms of the following keywords: cyanobacteria, blue-green algae, cyanotoxins, cyanobacterial blooms, harmful algal blooms, modeling, kinetic growth equations, mathematical models, prediction, forecasting, freshwater, etc. To produce a balanced and attainable literature search, a value of about 1,200 articles was arbitrarily chosen as the threshold. It was important to be specific with key words so that the search did not include hundreds of thousands of articles. However, the search was not so specific that it limited the results (e.g., <10 articles), excluding possibly beneficial articles. Several trial searches were completed using combinations of the key words given a high ranking. A comprehensive list of these search keywords, that were used in several combinations to compile articles, is provided in Appendix A.

While performing the literature search, articles were, to the extent possible, limited to areas with ecosystems and climates similar to Utah. This meant removing any articles related to marine or tropical freshwater cyanobacteria. Utah has more of a problem with cyanobacteria in lakes and reservoirs, so articles on rivers and wetlands were also removed from the search. The focus was on articles having to do with modeling, but articles relating to cyanobacterial gene, DNA, or RNA manipulation were additionally removed as they were out of the scope of this project. After several cycles of refining keywords, and removing duplicate articles that surfaced, there were 1,150 articles (Appendix B) that were reviewed for inclusion.

With such a large number of articles to manage, the reference management software Mendeley (Elsevier 2020) was used to provide a searchable database and direct access to articles' abstracts and, often, the papers themselves. The title, author, and year of each article was recorded in an Excel database, and then abstracts were read and a topic, relevance rating, and a short summary were additionally included in the document. The relevance rating was used to further sort and organize the documents and the highly relevant articles were further classified.

The rating of all 1,150 articles was not completed by two individuals, so to ensure the process was repeatable and to estimate scoring uncertainty, 20 articles were randomly chosen for quality assurance and scored by each individual, and the result classifications were statistically compared. Twenty articles were chosen as an arbitrary value to estimate rating variability without becoming overly time consuming. From the process, eight out of the combined 40 were given a different rating than the initial rating (e.g., rated “highly relevant” first and then being rated “moderately relevant” during the second rating). This repeat rating gave insight into whether the articles were being rated consistently throughout the process, and in hindsight, this process should have been done immediately after the first rating as well as at the end. Several rating checks would have been beneficial because the repeated rating was completed after the articles were read and reduced to the final number, and this impacted the results, showing seven of the eight articles being given a lower rank based on hindsight ranking knowledge. This process found that there was a lot of subjectivity introduced through the ranking processes. All in all, the repeated ranking showed that 80% of the articles were given the same rank as the initial rank.

The first round of ranking was highly inclusive leaving 439 articles rated as highly relevant. The abstracts of those articles were again read and only articles that discussed modeling were ranked as highly relevant. Articles that discussed types of cyanobacteria, cyanobacteria grown in the laboratory, or various methods to suppress cyanobacteria are examples of articles that were removed at this level. Sixty-four articles remained and were read in full, and articles were removed at this level based on earlier established criteria that were not listed in the abstract, for example models created for an estuary or a river, or if they had been previously covered in the Guven and Howard (2006) review. Ultimately this process condensed the database down to 34 highly relevant articles that would benefit and help build this report.

In accordance with the objective of this study, the relevant papers were examined in detail, and the models to predict HABs were summarized in this report (Appendix B).

Results

This report reviews a range of existing models produced to predict cyanobacterial bloom occurrences in lake and reservoir systems. The review contains the cyanobacterial prediction models published in peer reviewed articles from 2003 to 2020 selected by the methods described above. The models are grouped by (1) model methodologies, (2) prediction target, and (3) model specification, as shown in Table 4. The first category is based model methodology: mechanistic models or data driven models. Then they are separated based on upon what the models are used to predict: cyanobacterial biomass or cyanotoxins. The data-driven models are then broken into two further categories to describe the modeling approach: empirical and machine learning, the second of which further includes the categories: artificial neural network (ANN), support vector machine (SVM), boosted regression tree (BRT), hybrid evolutionary algorithm (HEA), multivariate adaptive regression spline (MARS), and other machine learning methods.

Many of the models, optimization techniques, and statistics are referred to by acronyms, and so a list of every acronym used in this report is found in the appendix (Appendix C). The author(s) of each article is (are) listed and then the developed model and results are summarized below each listed author(s). Each summarized section remains true to the article's layout and terminology in terms of name of model, reported errors, reported variance, etc.

Table 4. Classification of Cyanobacterial Models

Mechanistic Models	**Cyanobacterial Biomass**			Ibelings, et al. 2003 Wang, et al. 2012 Wang, et al. 2016 Fadel, et al. 2017 Vinçon-Leite, et al. 2017 Wang, et al. 2020
Data-Driven Models	**Cyanobacterial Biomass**	**Empirical**		Onderka 2007 Beaulieu, et al. 2014 Zhao, et al. 2019 Xu, et al. 2020
		Machine Learning	**Artificial Neural Network**	Teles and Vasconcelos 2006 Wang, et al. 2010 Ahn, et al. 2011 Millie, et al. 2014 Lou, et al. 2016 Luo, et al. 2017 Xiao, et al. 2017
			Support Vector Machine	Xie, et al. 2012
			Boosted Regression Tree	Liu, et al. 2020
			Hybrid Evolutionary Algorithm	Zhang, et al. 2015 Ostfeld, et al. 2015 Swanepoel, et al. 2016
			Other	Harris and Graham 2017 Mellios, et al. 2020
	Cyanotoxins	**Empirical**		Tao, et al. 2012 Hayes and Vanni 2018
		Machine Learning	**Support Vector Machine**	Alonso Fernández, et al. 2013 García Nieto, et al. 2015 García Nieto, et al. 2015 García Nieto, et al. 2017
			Boosted Regression Tree	Taranu, et al. 2017
			Multivariate Adaptive Regression Spline	García Nieto, et al. 2011 García Nieto, et al. 2012 García Nieto, et al. 2014

Mechanistic Models

A mechanistic model is based on the fundamental laws of physical, biological, and chemical processes in nature. Though more data is always better than less, this type of model requires a fraction of the experimental data compared with data-driven models (below) to estimate parameters and build scenarios. The parameters that are estimated (e.g., rate constants, partition coefficients, diffusivities) have actual scientific meaning, rather than a simple fit to experimental data.

Biomass

(Ibelings, et al. 2003)

Ibelings et al. (2003) began their model for cyanobacterial biomass with the analytical approach of a mechanistic model, but then used that framework to develop a fuzzy model (Von Altrock 1995). They chose to deviate from a fully mechanistic model because they require information that is often not available and is more prone to error propagation. Fuzzy logic is a system that does not use quantities but rather uses qualifications such as “high,” “intermediate,” or “low” to weight variables, which allows for working with uncertainties in quantitative inputs. All of the model inputs are translated to qualitative terms through a process called fuzzification to determine a qualitative prediction of the appearance of an algal bloom. Then, through a process classed defuzzification, the qualitative prediction is transformed into a relative numerical representation of the degree of surface bloom appearance.

This model was used to predict surface algal blooms on Lake IJsselmeer, a large lake in the Netherlands. The first (of three) component of the model was mechanistic and used differential growth rate equations based on grazing rates, sediment phosphorus release, and dispersive transport rates as well as hydrodynamics of the lake to calculate the cyanobacterial biomass in the lake. The second was to model the buoyancy and water column stability using fuzzy logic and inputting algal biomass and the other meteorological data variables into the model through fuzzification. Within the model, a series of “IF” and “THEN” rules were then applied to reach qualitative conclusions that then go through defuzzification to be converted to quantitative but relative values. The model-produced values predicted the appearance or disappearance of a bloom on an hour by hour basis. This model was extended with a third component to include a simple particle-tracking model, that uses transport differential rate equations, to be better applied in a spatial context, and the model was validated using satellite data from Lake IJsselmeer. Time lag is often a problem for both remote sensing and in-situ monitoring, so long-term weather forecasts were used to predict blooms to circumvent this issue.

The mechanistic model produced concentrations for cyanobacterial biomass that correlated well with field measurements. The model-predicted surface blooms overall matched well against satellite images of cyanobacterial blooms for the year tested. When applied to several years of data, the model predicted 270 of 290 cases correctly with 93% accuracy, but it failed to predict four surface blooms and then predicted 20 extra blooms that had no satellite image verification. Using a 95 percent confidence interval, it was found that the model was producing results that were more accurate than "by chance" predictions. The paper also emphasizes that the model often forecasts additional blooms that did not occur but rarely missed a bloom that did.

(Wang, et al. 2012)

Wang et al. (2012) developed a purely mechanistic model to predict blooms. A mechanistic model was chosen due to its simplicity and reliability. The model is a niche model (a form of species distribution models from ecology) based on taxonomy that calculates two phytoplankton community decisive variables (niche breadth and niche overlap) in space and time.

Understanding community structure, the theory of community succession, and resource competition leads to the prediction of growth and decline of cyanobacterial blooms.

This model was used to forecast cyanobacterial blooms in Chaohu Lake, a lake in China with serious eutrophication problems and yearly occurrences of cyanobacterial blooms. Eight sample points were chosen at which eight physiochemical factors were collected as well as a sample used to identify the most common species of cyanobacteria. Calculations were made based on the most common phytoplankton species. Niche breadth and depth were calculated and then a redundancy analysis (a variant of principal components analysis) software was used to determine correlations between environmental factors and cyanobacterial species. *Microcystis* was the dominant taxon that had the highest niche breadth in July, then other species such as *Anabaena* increased slightly in August, and finally other algae species increased in September. Then, every taxon started to decrease in October. Calculations for niche overlap and resource availability were performed to find a species growth rate and explain increasing and decreasing growth of each species.

The redundancy analysis revealed that water temperature, dissolved oxygen, and total dissolved phosphorus were the most significant environmental factors that affect the dominant *Microcystis*. The model could forecast species potential growth rates. The model-predicted growth rate and actual biomass had a correlation coefficient of 0.988, and thus the model successfully predicted the growth and decline of cyanobacterial blooms.

(Wang, et al. 2016)

Wang et al. (2016) developed a mechanistic integrated hydro-environmental model. The model was developed to provide a *Microcystis* bloom warning for Jinshan Lake, a lake in China that has been experiencing *Microcystis* blooms since infrastructure was implemented to control water exchange processes. Microcystis blooms are influenced by the combined effects of environmental factors and nutritional factors, so a bloom driving function was derived. The function includes three subfunctions: the first directly discriminates the area where a bloom does not occur through an environmental restriction function, the second and third are used to determine where a bloom may easily occur through an exponential function of water nutritional status and an integral contribution function of environmental factors' effects on *Microcystis* bloom activation energy, respectively.

The bloom driving function is the basis for the model. The model is composed of five modules: input, simulation, solution, discrimination, and output. In the input module physical, chemical, and biological boundary conditions are established. In the numerical simulation module, six processes (light intensity, nutrient concentration, temperature, biological inhibition, water current, and Chl-a concentration) are simulated using differential mass and energy balances, and kinetic equations. The function solution module is where the function described above is applied, followed by a discrimination module that produces detailed bloom warning grades and then an output module that relays what the level of risk of HABs would be given certain conditions. The lake was split into five regions and then the model was run to indicate what conditions would cause a bloom in each region. One region, based on its characteristics, was identified as the location with the highest bloom risk or where blooms would occur first.

(Fadel, et al. 2017)

Fadel et al. (2017) developed a simple one-dimensional ecosystem model, coupled with a hydrodynamic model to simulate the succession of cyanobacterial biomass with correct magnitude and timing. The model is referred to as a dynamic reservoir simulation model – computational aquatic ecosystem dynamics model (DYRESM-CAEDYM) (Imberger and Patterson 1981). The DYRESM part of the model is able to simulate the vertical distribution of temperature, salinity, and density in a water column, while the CAEDYM part of the model can simulate nutrient cycles and zooplankton and phytoplankton dynamics. This model was created to describe conditions in Karaoun Reservoir, a waterbody used for hydropower and irrigation in Lebanon. Data collected for this reservoir included hydrologic, meteorological, and physical, chemical, and biological field measurement data.

The DYRESM-CAEDYM was configured to simulate changes in water level, temperature, and cyanobacterial biomass through hydrodynamic and kinetic equations. The model was calibrated using 1 year of measurements and then verified using 2 years of measurements for water level and a separate year of measurements for water temperature. The simulated and observed data were very comparable. The root mean square error for each model verification was less than 1 m or less than 1 °C.

Two species were predicted in the model: *Microcystis* and *Aphanizomenon*. According to the model, the *Aphanizomenon* was quickly increasing in May and June, not limited by temperature and moderately limited by light. After a peak in mid-June it began to decrease, being limited by temperature and occasionally limited by light. The *Microcystis* started increasing the beginning of June and increased until it peaked the end of August, not being limited by temperature and moderately limited by light. The model was better able to predict *Microcystis* biomass with coefficients of determination greater than 0.85 for 2012 data, but was less successful in predicting *Aphanizomenon* with the greatest coefficient of determination being 0.66.

(Vinçon-Leite, et al. 2017)

Vinçon-Leite et al. (2017) developed a one-dimensional vertical, physical-ecological model for cyanobacterial growth and provide an early bloom warning system. This model uses a mechanistic approach to simulate cyanobacteria growth and nutrient dynamics. The study site for this model was the Yuqiao Reservoir, an important drinking water source in China. A historical data set of physiochemical and biological data was used for model calibration.

The model calibration consisted of using 7 days of values for calibration and the next consecutive 11 days for validation. The model was calibrated by adjusting parameter values to minimize the root-mean square error between simulated and observed results. Calibration was performed with each of the parameters used to produce the output of predicted temperature and cyanobacterial biomass represented as mg/L Chl-a. The calibration results showed correlation coefficients of 0.91 and 0.62 for temperature and cyanobacterial biomass, respectively. The validation results showed correlation coefficients of 0.62 and 0.48 for temperature and cyanobacterial biomass, respectively. In the end, the model used inputs of cyanobacterial biomass and temperature for initial conditions and used meteorological forecasts to predict cyanobacterial biomass for the next five days.

(Wang, et al. 2020)

Wang et al. (2020) developed a nonlinear mathematical model of cyanobacteria growth to forecast bloom outbreaks. This model is the combination of a mechanism-driven model and a data-driven model to overcome the shortcomings (data-driven models only carrying out simple data analysis because of the lack of theoretical support, and mechanistic models being very complex, and involving many parameters that can introduce uncertainty) of each model type. First, the cyanobacteria's demand for phosphorus and nitrogen was calculated using growth stoichiometry. Then the mathematical model was developed to simulate cyanobacterial growth using differential growth kinetics equations and four main factors (Chl-a concentration, temperature, nitrogen concentration, and phosphorus concentration).

The data-driven part of the model was developed to estimate and optimize parameters for the non-linear dynamic growth model. Historical environmental data were used to calibrate the model using the Cuckoo algorithm (Yang and Deb 2010), which is an iterative optimization algorithm based on the behavior of the nest stealing cuckoo bird. A bifurcation set of equations, using the established parameters and based on the cusp catastrophe theory (Zeeman 1976), was then used to determine the critical point of a cyanobacterial bloom outbreak. If the final calculated value of the bifurcation set within the nonlinear model is approximately zero, it indicates that the established environmental parameters have reached a critical point for cyanobacterial bloom outbreak, thus forecasting or creating a cyanobacterial warning system.

Data-Driven Models

Data-driven models often take years of experimental data to build. Two forms of data-driven models are discussed in this section: traditional empirical models and machine learning models. Empirical model building tries to find the simplest mathematical explanation for the features of the data set without resorting to scientific mechanisms. The fit they create to the existing data is used to provide future estimates. Machine learning models use large amounts of data for training information to discover relationships between inputs and outputs that improve the model's mathematical algorithms as more input data is added to predict future values.

Empirical Models

Empirical data-driven models are based on evidence or observation rather than theory. Equations are fitted to a proposed curve or line to predict future values in an empirical model. Empirical models can also include general linear models that compare how several variables effect a response variable, generalized linear models, which is an extension of the general linear model but allows for non-normal response variables, and generalized additive models (GAMs). A GAM (Hastie and Tibshirani 1990) is an extension of a generalized linear model, meaning that it is a non-parametric generalization of multiple linear regression. A GAM is more advanced (using additive regression) than simple linear regression models but is not as sophisticated as many machine learning models. A GAM is useful for modeling many complex environmental systems because it does not require normally distributed data and can incorporate non-linear relationships between environmental factors and responses, similar to ANN models. The additive functions within a GAM allow for partial effects or contributions of each independent variable to be realistically obtained.

Biomass

(Onderka 2007)

Onderka (2007) developed a simple regression predictive model to predict bloom risk. This model uses six years of data for phosphorus, nitrogen, and water temperature levels in Liptovska Mara Reservoir, a eutrophic reservoir in Slovakia used for recreation and to produce hydropower. The data values of total nitrogen, phosphorous measured as phosphate, and temperature were graphed against cyanobacterial cell counts measured as chlorophyll a (Chl-a).

Three regression models were computed to develop three empirical equations that describe the data. The three equations used phosphate, phosphate and temperature, and phosphate and total nitrogen. The phosphate and total nitrogen relationship showed the best results to predict cyanobacterial blooms although no model accurately predicted the magnitude of the peaks. Ultimately the model revealed that blooms developed during periods of increasing N:P ratio, and the model could be used as a simple way to predict future bloom occurrences by measuring three parameters.

(Beaulieu, et al. 2014)

Beaulieu et al. (2014) applied linear and nonlinear models to 149 lakes to determine whether empirical cyanobacterial prediction models vary regionally. The models were constructed to assess the relative importance of nutrients, water column stability, and water temperature in driving cyanobacterial biomass. Data were collected from waterbodies across three provinces of Canada, and a year of data was used and condensed to the growing season.

The paper considered general linear models and nonlinear GAMs to predict cyanobacterial biomass. A mixed-effects model was additionally produced to introduce the different provinces as a random effect. Model selection was determined by an analysis of variance and Bayesian information criterion (BIC). Across all the lakes, nutrient concentrations were found to be strong predictors of cyanobacterial biomass, and total phosphorus or total nitrogen performed equally well through linear modeling. pH was also determined to be a significant factor in prediction, but N:P ratio was determined to be insignificant. The linear models outperformed the nonlinear GAMs based on the BIC. The mixed-effects model also showed that there were no significant regional differences in cyanobacterial biomass response models. Overall, the paper concludes that the models they produced can be used to estimate cyanobacterial biomass from any Canadian region.

(Zhao, et al. 2019)

Zhao et al. (2019) developed a probability prediction model based on cyanobacteria species and cyanobacteria driving factors to predict the occurrence of cyanobacterial blooms. Physical, chemical, and biological data collected for this model were from 13 reservoirs and lakes in Jinan City, China. First, dominant cyanobacterial species were identified using a dominance model (Zhao, et al. 2014) and the mutation point method (Zhao, et al 2015), and then determination of principal cyanobacterial driving factors was completed using a canonical correspondence analysis (CCA). A CCA determined relationships between biological groups of species and water quality factors through a multivariate gradient analysis.

The probability model included six species of cyanobacteria that were determined to be dominant as well as two physical water quality factors (water temperature and pH) and four chemical water quality factors (total phosphorus, ammonia nitrogen, chemical oxygen demand, and dissolved oxygen) that were determined to be principal cyanobacteria bloom driving factors. The model calculated a probability of the occurrence of cyanobacterial blooms at each monitoring station and established that if the probability was greater than 0.75, all of the driving factors were within their optimal range, and the risk of a bloom was high.

(Xu, et al. 2020)

Xu et al. (2020) developed a GAM to identify the effects of nutrients on algal biomass to predict HABs. A GAM was used because a nonlinear framework is needed for modeling complicated nutrient and algal biomass relationships. This GAM incorporates external nutrient loading to the lake and a cumulative loading term with different time scales to account for nutrient recycling from lake sediments, which can often fuel blooms even when external loading is decreased. The model uses data from Lake Okeechobee, a lake in Florida that is seasonally dominated by HABs. Chl-a is used as the response variable to identify algal biomass and nutrient concentrations are used as inputs.

This study identifies the relative importance of each variable by excluding it from the GAM and then focusing on the model’s goodness of fit when the variable is dropped. It was found that better results were determined when the cumulative loading term was added and not just external loading inputs. The paper determined that there was an inhibiting effect at high total phosphorus levels and that the lake was nitrogen-limited. It concluded that reducing phosphorus levels at certain times would actually increase Chl-a concentration. The best model explained 52.8 to 72.8% of Chl-a concentration. The paper concludes that both nitrogen and phosphorus are important in modeling blooms and that spatial heterogeneity should also be considered.

Cyanotoxins

(Tao, et al. 2012)

Tao et al. (2012) developed a GAM that uses abiotic and biotic factors to predict toxic cyanobacterial blooms. A GAM was chosen due to its ability to incorporate non-linear relationships between input variables and responses as well as its ability to handle non-normally distributed variables. The model was developed based on one year of data collected from Lake Taihu in China, a drinking water supply for millions of people that is seasonally dominated by toxic blooms that are mainly composed of *Microcystis*.

After data collection, a statistical regression analysis was performed in the GAM to assess the effect of the environmental factors on microcystin production during growth of a bloom. The model found that microcystin concentration was best explained using the factors conductivity, dissolved inorganic carbon (DIC), water temperature, and pH, but additionally found that neither nitrogen nor phosphorus had significant correlation with microcystin production. The model found that *Microcystis* abundance did not always correlate with microcystin concentration. Ultimately the model found that conductivity, DIC, water temperature, and pH could, respectively, explain 21, 12, 11, and 10 percent of the variance related to microcystin concentrations. The study concludes that this model could be used to predict toxic blooms and microcystin concentrations using the correlation of four important parameters.

(Hayes and Vanni 2018)

Hayes and Vanni (2018) developed several regression models for predicting microcystins (hepatotoxins associated mainly with *Microcystis*). The purpose of each model was to use watershed characteristics, lake morphometry, nutrient concentrations, and biological characteristic to predict microcystin concentrations. Models were produced using data from 136 lakes and reservoirs in Ohio to determine the ability for parameters such as lake morphometry, nutrient concentrations, biological characteristics, and watershed characteristics to predict microcystin concentrations. The data were collected in 2006, 2007, and 2009. An analysis was performed on the full data set, and then on only the 2009 data.

Before inputting data into the model, many of the inputs were transformed to reduce skewness and a censored regression analysis was performed to account for the large number of microcystin observations that were below detection limits. Corrected Akaike information criteria (AICc), that estimates out-of-sample prediction error, was used to select the best-fit model.

The models identified that microcystins were positively correlated with Chl-a, TN, TN:TP, and percent agriculture in the watershed and were negatively correlated with the ratio of watershed area to lake surface area (WA:SA). The equation selected for the inclusive data that was the most parsimonious with the lowest AICc value included both WA:SA and Chl-a predictor variables. For the 2009 data, two equivalent equations were chosen, one including biomass of potentially toxin producing cyanobacteria and the other including Chl-a. The study concluded that elevated microcystin levels were likely to be found in lakes with smaller WA:SA and higher biomass concentrations.

Machine Learning

Machine learning models use computer algorithms to learn and improve from input data. These models include Artificial Neural Networks (the most common), Support Vector Machines, Multivariate Adaptive Regression Splines, Hybrid Evolutionary Algorithms, and others. The language used to describe each model, including error reporting, accuracy reporting, terminology of model inputs, etc., is true to the language from each paper.

Artificial Neural Network (ANN)

ANN models are machine learning models that can be supervised, meaning they learn from data known to be correct, or unsupervised models that work on their own to discover connections. Most of the models discussed in this section are supervised models, in that they are provided with environmental conditions and their related result on cyanobacteria.

Supervised ANN models consist of a so-called multilayer perceptron (MLP) design that includes an input layer and an output layer created by the data provided, and between these two layers are hidden layers that use an error convergence technique to determine weighted connections between inputs and outputs. The hidden layers consist of nonlinear functions (generally sigmoid or hyperbolic tangent functions) that create interconnections between neurons and adjust to decrease the output error. These models often use back propagation that allow errors to be reduced when an error function is propagated back to the hidden layer to update the weights of each connection.

A self-organizing map (SOM) is an unsupervised learning machine that consists of two layers, an input and output layer of neurons. It produces a map or two-dimensional grid through competitive learning rather than error reduction in supervised learning. The learning algorithms produce different results in response to input patterns.

Biomass

(Teles and Vasconcelos 2006)

Teles and Vasconcelos (2006) produced an ANN model that uses nonparametric, data-driven, self-adaptive methods and requires very few theoretical assumptions. An ANN model was used because the prediction of population growth of cyanobacteria was more important to the study than the identification of underlying environmental processes or kinetics affecting the growth. The specific type of ANN they used was a generalized regression neural network (GRNN) (Specht 1991) and was used to predict cyanobacteria abundance in Crestuma Reservoir (Douro River, Portugal), a reservoir periodically dominated by toxic *Microcystis* and *Aphanizomenon* taxa, and used for recreation, drinking water, and power production.

Physical, chemical, and biological data collected from the reservoir over a three-year time period were split and one portion was used for training while the other was used for verification. Input parameters for the model were carefully selected, using a cluster analysis and model sensitivity analysis, to prevent overfitting, which is beneficial for model training but not particularly useful for forecasting. An appropriate time lag was carefully selected by running the GRNN and selecting the time lag that provided the best correlation between input and output values. A

GRNN was selected by using combinations of variables in a forward and backward stepwise manner until the best correlation was observed.

Ultimately, after creating and training the models, the remaining data were used to verify each model's ability to predict cyanobacterial biomass in units of cells/mL. Models were run with the selected 15-day time lag, and were run using physical and chemical inputs, phytoplankton taxa inputs, and all variable inputs. After verifying the models, independent observations from 1999 were used for final validation. The highest correlation coefficient for the verification test (r = 0.878) was found with the model using only physical and chemical input parameters, and the highest correlation coefficient for the validation test (r = 0.773) was found using all input variables.

(Wang, et al. 2010)

Wang et al. (2010) developed a model that uses a back-propagation neural network (BP-ANN) and rough set decision method (Greco et al. 2001) to predict cyanobacterial blooms. An ANN model was selected because it could combine complex physical, chemical, and biological inputs and simulates cyanobacterial growth using nonlinear, machine learning processes with high prediction accuracy. The ANN was paired with a rough set decision theory because it is a useful tool for data mining that removes irrelevant parameters without losing useful information. A ChiMerge algorithm was used in the rough set decision theory to convert continuous values into interval values to determine which factors would remain in the model and to generate decision rules. The rough decision rules were incorporated into the model to form a hybrid (BPANNRD) and the original BP-ANN model was run in parallel to determine if incorporation rough decision rules improved the model.

The model used five years of data (April to October) from Dianchi Lake, a lake affected by toxic *Microcystis* taxa and used for recreation and agricultural activities. Daily weather data were collected including average, highest, and lowest temperatures, precipitation, sunshine duration, average and highest wind velocity, and wind direction. They also used cyanobacteria remote sensing data and paired the two data groups day by day. The data were separated into two parts to train and then verify the model.

The results confirmed that combining the BP-ANN with a rough set decision theory increased its testing accuracy by 3 percent in training and 3.6 percent in verification when compared to the BP-ANN run in parallel. Ultimately, the BPANNRD had an 84.3 percent accuracy at predicting the occurrence of a bloom using only weather and remote sensing data. It was concluded that the rough set decision allowed for useful parameter selection and improved the model, which could be used to forecast algal blooms.

(Ahn, et al. 2011)

Ahn et al. (2011) developed two ANN predictive models to predict cyanobacterial density. ANN models were chosen because they do not demand that there must be linear relationships between input variables and are efficient at dealing with systems that have complex nonlinear relationships. The ANN models in this study were a SOM, that efficiently visualized relationships between cyanobacteria and environmental factors and traced temporal change patterns in environmental conditions, and an MLP, that used the environmental factors to provide good predictive power.

The models were created using five years of chemical and biological data (spring to autumn) from Daechung Reservoir. Fourteen independent environmental variables were compared to the dependent variable of cyanobacteria density using a cross-correlation function. Using this process, a time lag with the highest associated correlation was chosen for the models. The SOM consists of two layers, an input and output layer of neurons. The input layers of environmental factors and cyanobacteria density formed an output layer array of 54 outputs in a hexagonal lattice. A connection intensity or weight between the input and output layers is created by the SOM algorithm and then the weight is consistently updated by SOM learning rules until the model is optimized and an output is chosen that has a weight with the least distance from the input.

The MLP uses a backpropagation algorithm designed to minimize the mean square error between computed and desired outputs. For the MLP they used all the samples except one to train the model and then used one sample to verify the model. The training process was performed until the sum of squares error was less than 0.001 in predicting cyanobacterial biomass (cells/mL) to avoid overfitting.

Cross-correlation found that water temperature was the most correlated with future cyanobacterial density, and the highest correlation happened at 2-4 weeks lag time, and linear regression estimated that cyanobacterial density was best predicted with a 3-week lag. The SOM model provided connection intensities in the form of a lookup table where the cyanobacterial density could be found by matching a cell with the most similar environmental conditions. The results showed relatively low predictability, with a low coefficient of determination, especially for bloom peaks. The SOM could be useful for monitoring changes in environmental conditions but was not recommended for prediction. The MLP model, on the other hand, showed a very high coefficient of determination when comparing the predicted and observed data. Using a sensitivity analysis within the MLP, it was confirmed that water temperature was the most contributing variable and that dissolved nitrogen compounds also played a large role, whereas phosphorus did not. The MLP was recommended to predict a cyanobacterial bloom 3 weeks in advance.

(Millie, et al. 2014)

Millie et al. (2014) developed an ANN to model total phytoplankton and *Microcystis* biomass. An ANN model was chosen because environmental conditions are often chaotic and the model allows nonlinear patterns to be identified and reproduced and provides impartial predictions so that patterns between environmental conditions and biomass can be discovered. The ANN model was created using 3 years of data from western Lake Erie (USA and Canada), a lake plagued with toxic populations of *Microcystis* seasonally. Data collected consisted of abiotic, meteorological, and biotic variables.

Before the model, a coefficient of variation was calculated to assess the uncertainty in the abiotic and metrological variables, while an analysis of variance was performed on the biotic variables to assess annual and monthly difference of Chl-a concentrations. A connected weight and global sensitivity analysis was used to identify the relative importance of each input variable. The analysis identified water temperature and wind speed as the two most influential parameters. The model is an MLP design using a backpropagation algorithm that has three layers including an input layer of significant variables, a hidden layer that uses hyperbolic tangent transfer functions and a momentum learning algorithm, and an output layer that provides information for future predictions.

The model was run to identify phytoplankton biomass via Chl-a concentrations and then to identify *Microcystis* biomass. A subset of the data was used to train the model, and then provide cross-validation before the final test. For the phytoplankton biomass, the model identified total phosphorus as a variable that has the greatest predictive influence. However, for the *Microcystis* biomass, all variables were equally predictive and no single variable greatly influenced the prediction. The models had correspondences (a discrete form of correlation, based on classification) of 0.87 to 0.97 and 0.70 to 0.94 for phytoplankton and *Microcystis* biovolumes, respectively.

(Lou, et al. 2016)

Lou et al. (2016) used the data from the Macau Reservoir SVM model (Xie, et al. 2012) to build an extreme learning machine (ELM). An ELM model (Guang-Bin, et al. 2006) is a single-hidden-layer, feed-forward neural network that can solve nonlinear and complex problems. This study produced both a prediction and forecast model and found better results than Xie’s (Xie et al. 2012) SVM study, with coefficients of determination of 0.83 and 0.90 for the prediction and forecast models, respectively.

(Luo, et al. 2017)

Luo et al. (2017) developed a combination prediction model of a back-propagation ANN (BPANN) and an adaptive grey model (AGM). An ANN model was chosen because of the complexity of the system and the nonlinearity between variables. The data collected for this study was collected using remote sensing as it can offer unattended, continuous, and long-term observations over a large area. Multi-source heterogeneous water environment sensors were used to gather continuous and precise data that was applied to the model.

The data for this model were retrieved from Dianchi Lake, a Chinese lake with large environmental and aesthetic importance. To allow for enough physical, chemical, and biological indicators to provide accurate predictions without over-fitting the data with too many input variables, the researchers selected 13 parameters they deemed as highly related to eutrophication. The model used these indicators to forecast the Chl-a concentration within the next 24 to 72 hours.

The AGM part of the model, used to fill in data gaps in the model, evolves from a first-order differential equation and is based on an exponential time series function. It is able to take random, incomplete, and uncertain data and describe future tendencies of water quality data based on previous or incorrect information. The outputs for the AGM model were used to train the BPANN model. The BPANN is a model that can solve high-dimensional, nonlinear problems and can transform data into something that is definite and reliable. The model has an MLP design, uses forward propagation, and the hidden layer consists of sigmoid or hyperbolic tangent functions.

The AGM model showed good water quality indicator prediction performance with coefficients of determination ranging from 0.81 to 0.88 for the parameters. The BPANN model was successful in predicting Chl-a concentrations with a coefficient of determination of 0.93. The paper concludes that the model is a successful 1-day in advance early warning system for cyanobacterial outbreaks.

(Xiao et al. 2017)

Xiao et al. (2017) developed a combination model of a wavelet analysis with an artificial neural network (WNN). An ANN model was chosen because of the complex physical, chemical, and biological parameters that are involved in the system, and is able to handle environmental non-linearities. ANN models often have difficulty dealing with non-stationary data and so a wavelet analysis (Lee and Yamamoto 1994) was used in combination as an effective tool for revealing both spectral and temporal information simultaneously. This combination model uses a single parameter (cyanobacterial biomass as Chl-a concentration or cell density) as the key input, predicting future cyanobacterial biomass from previous biomass.

This model was tested in two water bodies. The first set of data to build this model was collected from the Siling Reservoir in China, that provides drinking water, irrigation, recreation, and power generation, yet consistently suffers from HABs. Real-time biological data of cyanobacterial density and Chl-a concentration were collected from a buoy monitoring station, using a HACH Hydrolab DS5 with multiple probes, as daily average values. The second data set came from Lake Winnebago, a large calcareous lake in Wisconsin that provides both recreation and drinking water beneficial uses. Chl-a concentrations were collected from a buoy in the form of daily averages.

The hybrid model was created by, first, transforming the original time series into time-frequency domains by a discrete wavelet transformation process. Then each wavelet decomposition series was used as an input to each back propagation artificial neural network. This ANN model is an MLP design and uses back propagation. The input data were split 60% for training the model, 20% for validation of the model, and 20% for testing the successfulness of the model.

The models produced wavelet decomposed forecasting series which were summed to forecast an algal cell density series. For forecasting cyanobacterial density in Siling Reservoir the model had good results with the best model having a correlation coefficient of 0.986 for 1 day early cyanobacterial cell density predictions. For forecasting Chl-a concentrations in Siling Reservoir, there was good predictive power with a correlation coefficient of 0.900 and for Lake Winnebago a correlation coefficient of 0.872. Both model results indicated that the WNN could be used to reliably predict cyanobacterial blooms.

Support Vector Machine (SVM)

An SVM model is a machine learning technology that is based on statistical theory and can solve complex nonlinear problems (Vapnik 1995). SVM is derived from instruction risk minimization, which allows it to minimize generalization error (Xie, et al. 2012). The model has an input layer of environmental parameters, hidden layers that use quadratic programming and kernel functions, and an output layer that provides information for predictions. SVMs were originally developed for classification but were later generalized to solve regression problems through a method called support vector regression (SVR) (Alonso Fernández, et al. 2013).

Biomass

(Xie, et al. 2012)

Xie et al. (2012) developed an SVM model to predict dynamic change of algae and cyanobacteria populations in water. An SVM model was chosen because it has been shown to solve complex nonlinear environmental problems with high prediction accuracy over long prediction periods while only requiring a small number of samples. Two models were created in this study: a prediction model that does not consider time series effects, and a forecasting model that does consider time series effects.

The models were created using data from Macau Storage Reservoir, a reservoir in China that provides drinking water and suffers from HABs. Eight years of physical, chemical, and meteorological data were used to train the model and three years were used to test the model. A correlation analysis was performed to determine which parameters were significantly correlated with biomass. The SVM model was built using a radial basis function (kernel function) and was built to have internal cross-validation to avoid overfitting. The prediction model had nine parameter inputs and the forecast model had 23 time-lagged parameter inputs. Ultimately, both models performed well with square of correlation coefficients of 0.760 and 0.863 for the prediction model and forecast model, respectively. The paper recommends the SVM model as an effective approach for predicting blooms.

Cyanotoxins

(Alonso Fernández, et al. 2013)

Alonso Fernández et al. (2013) developed an SVM model to predict cyanotoxin presence. The SVR method was chosen within the SVM model because it is a non-linear generalized regression method that has a flexible procedure to model complex variable relationships without strong assumptions. The SVR also requires a shorter training time when compared to other machine learning techniques. The model was created using five years of biological and physicochemical data from the Trasona Reservoir, a reservoir in Spain used for recreation and industrial water supply. The goal of the model was to identity the dependence relationships between the cyanotoxins and collected input variables.

This SVR model uses kernel parameters to map nonlinearly separable data into a feature space where it is linearly separable. The model was run using linear, polynomial, radial basis, and sigmoidal kernel functions. The sigmoidal kernel functions produced the best model with correlation coefficient of 0.93 when comparing predicted toxin concentrations to observations. To guarantee predictability a cross-validation algorithm was used to split training data into three groups: a training dataset (80%), a validation dataset (10%), and a testing dataset (10%).

(García Nieto, et al. 2015, García Nieto, et al. 2015, and García Nieto, et al. 2017)

The following three articles used the same data for the Trasona Reservoir and base SVM model (Alonso Fernández, et al. 2013) but used different optimization techniques to find input variables that lead to the best prediction results or use different kernel function in the model. García Nieto et al. (2015) used the SVM model in combination with a particle swarm optimization (PSO) technique (Olsson 2011) to forecast the presence of cyanotoxins. An SVM model was chosen due its well-known ability to predict values from very different fields and approximate any multivariate function to any desired degree of accuracy. The PSO technique is an evolutionary computation, swarm intelligence-based algorithm used to optimize parameter selection for the SVM model.

The PSO uses bio-inspired algorithms (Kennedy and Eberhard 1995) to iterate proposed solutions (or parameters), evolving with each iteration based on individual and neighboring trajectories, toward an optimal solution. The PSO optimizes SVM parameter inputs by comparing the forecasting error in every iteration. A radial basis kernel function was chosen for the SVM model due to the function's effective approach to nonlinearities present in the regression problem. The input data set was split into 10 parts where nine parts were used for training and then one part was used for testing. To guarantee prediction ability, this process was performed 10 times using a 10-fold cross-validation algorithm. The model had a very good fit between the model and the observations with a correlation coefficient of 0.95. Overall, the addition of the PSO technique produced better results that the preceding model, and the improved model is recommended as an accurate predictor of cyanotoxin concentrations in this reservoir.

García Nieto and others (2015) used the SVM model in combination with an artificial bee colony (ABC) technique to predict cyanotoxins. An ABC is an evolutionary computation algorithm (Simon 2013) similar to PSO that optimized input parameters to the SVM model inspired by the intelligent foraging behavior of honeybee swarms.

The ABC technique shares information between individuals in the population to find an optimum strategy. The ABC has three phases that are repeated until a maximum number of iterations occur or the algorithm converges. A radial basis kernel function was chosen for the SVM model and the 10-fold cross-validation described above was used to ensure predictability, and the best combination of parameters was chosen. The model had a very good fit between model and observed data with a correlation coefficient of 0.95. The model was run again, ranking variables, and adding variables until the highest correspondence was found. Using fewer input parameters gave a correlation coefficient of 0.96, overall having better results that the SVM model alone, and slightly higher, but very similar results to the PSO combination model that used all of the parameters.

García Nieto and others (2017) modified the model to a wavelet kernel SVM model in combination with the ABC technique to identify and predict cyanotoxins. The SVM model and ABC technique were developed using the same data and approaches described above (Alonso Fernández, et al. 2013 & García Nieto, et al. 2015). This model introduced wavelet analysis in which a wavelet is a mathematical tool that can be used to extract information from different kinds of data by correlating data that contains information of similar frequencies. A wavelet kernel function was derived, specifically called the Mexican hat wavelet kernel, and was combined with the SVM model.

The model was run and validated using the 10-fold cross-validation describe above to ensure predictability. The Mexican hat kernel function created a more predictive model and was a better approach to the nonlinearities present in the regression problem than the other kernels tested, showing a correlation coefficient of 0.954 versus the previously described radial basis kernel approach with a correlation coefficient of 0.95.

Boosted Regression Tree (BRT)

A BRT model (Elith, et al. 2008) fits and combines many models for prediction using two algorithms: decision trees, which use sets of rules in a tree-like model and boosting, which constructs and combines a collection of models. BRT models build and merge results from multiple models using a stage-wise forward selection procedure that combines numerous simple regression trees to create an optimized predictive model (Taranu, et al. 2017).

Biomass

(Liu, et al. 2020)

Liu et al. (2020) developed a BRT model to predict cyanobacteria biomass. A BRT model was chosen because it can model nonlinear responses and interaction of variables and can model different effects of covariates. This model used diatom-inferred total phosphorus (DI-TP) to predict cyanobacterial biomass using data from around 1000 U.S. lakes. Before modeling, a general lake type classification including: deep, shallow, natural, and manmade was determined as well as a classification on cyanobacteria functional groups including: potential N_2-fixing, heterocyst-producing, potential toxigenic, and bloom-forming.

The National Lake Assessment (NLA) (EPA, 2009) provided data sets including water chemistry condition estimates, green algae and diatom count data, and sampled lake information. DI-TP concentration was calculated using relative abundances of diatom taxa in the surficial sediments of lakes and a weighted average model. DI-TP was used because it can more accurately assess bioavailable phosphorus than total phosphorus can. DI-TP along with eight other predictor variables were used in the model to predict cyanobacteria biomass. A function was applied to the model to select for the most influential predicator variables and then the model was run several times with different variable combinations.

Modeling results found that DI-TP was a good predictor variable across different lakes and different cyanobacterial function groups but was not as important as water column total phosphorus. DI-TP performed better as a prediction variable when modeling N_2-fixing cyanobacteria over other functional groups and also performed better in deep man-made lakes over other shallow or natural lakes.

Cyanotoxins

(Taranu, et al. 2017)

Taranu et al. (2017) developed a BRT model paired with a hierarchical zero-altered model to predict when and where cyanotoxins will be produced. The hierarchical model was used to determine the importance of environmental multi-scale interactions. It was used to separate out and remove non-detect microcystin concentrations and understand the environmental interactions driving microcystins above non-detect limits. The BRT model was used to identify what environmental thresholds are associated with microcystins causing severe impairment to a waterbody. The environmental parameters that were considered included physiochemical variables measured in the water column, site and catchment characteristics, and land-cover variables.

The models were created using a continental-scale data set of 1127 lakes, ponds, and reservoirs randomly selected across the United States. A large number of these waterbodies (68%) had microcystin concentrations below the detection limit, creating a largely right-skewed data distribution. Ignoring these below detection results could result in exaggerated estimates of variance and biased estimates of parameters and standard errors, so a hierarchical two-state model was designed. It was important that this model was a zero-altered hurdle model (Zuur and Ieno 2016) instead of a zero-inflated mixture model (Phang and Loh 2013) because the interest was in determining the probability of not measuring detectable microcystins versus measuring any quantity of microcystins. This model had a binomial generalized linear mixed model (GLMM) to determine the presence or absence of microcystins in the waterbodies, and a detection limit-truncated log-link Gamma model that removed all non-detect data and modelled the remaining data as a function of selected environmental variables to identify drivers after microcystin presence is detected.

To further identify predictors of when cyanotoxin concentrations would be above or below toxic level guidelines, a BRT model was developed using all of the data sets. The model is able to identify optimal parameters through its stage-wise process, and then to simplify results, explanatory variables are dropped until the predictive deviance exceeds the original standard error as a function of the number of variables removed.

The binomial GLMM explained 55% of variance for the presence-absence of microcystins, and occurrence was best explained by local and regional features of nitrogen enrichment, cyanobacterial biomass or Chl-a concentration, dissolved organic carbon, depth, latitude, longitude, and effect of ecoregion agriculture. The second part of the hierarchical model, that was restricted to sites where microcystins were detected, explained 26% of the variance and showed that only local features were identified as predictors including, total nitrogen, cyanobacterial biomass, and turbidity. It also showed that microcystins were more abundant in natural lakes than in reservoirs. The BRT model showed that cyanobacterial biomass and elevated nitrogen concentrations were associated with microcystin concentrations and little to no interaction was identified among environmental drivers.

Multivariate Adaptive Regression Spline (MARS)

MARS is a non-parametric regression technique that is able to automatically model complex, nonlinear problems and interactions between input factors (García Nieto, et al. 2011). A MARS model does not require theoretical assumptions about relationships between the independent and dependent variables. Instead the relationship is discovered by functions driven by regression data.

(García Nieto, et al. 2011)

García Nieto et al. (2011) developed a MARS model to identify cyanotoxin presence. A MARS model was chosen because it has a structure able to handle complex, nonlinear environmental problems. The data used to build this model was five years of biological data from Trasona reservoir, a recreation and industrial use source in Spain.

The MARS model uses a nonparametric regression technique to predict a dependent variable (cyanotoxins in mg/L) from a set of input explanatory independent variables. Eight prediction variables were used as inputs to the model. Predictability of the model was insured by using a cross-validation algorithm that created a model for each observation and used the remaining data to train the model. The model successfully produced prediction results with a correlation coefficient of 0.91 with the observations.

(García Nieto, et al. 2012 and García Nieto, et al. 2014)

The following two articles use the same base MARS model and improve model inputs to generate better output predictions. García Nieto et al. (2012) developed a MARS to improve the prediction accuracy of previous work (García Nieto, et al. 2011). The model improved the inputs by using six years of not only biological data, from Trasona reservoir, but physiochemical data as well. The same model building and validation procedures were followed. Adding physiochemical information improved the prediction results; the final MARS model had a correlation coefficient between the predictions and observations equal to 0.97.

García Nieto and others (2014) used the MARS model in combination with a genetic algorithm (GA) to forecast cyanotoxin presence. The GA was used to reduce the number of input predictor variables based on their importance to the system. The model was developed using six years of biological and physiochemical data from the Trasona reservoir, and then six years of samples were used from 15 reservoirs in the same area to determine the model's ability to accurately predict cyanotoxin risk within the entire study area.

GAs have proven to be a powerful and reliable optimization technique. The algorithms are based on an interpretation of the evolutionary process, and input variables are chosen based on a fitness value, only allowing the strongest to survive. The GA process optimized input parameters, reducing the input from 24 to 6, and created a more successful model than previous MARS models with a coefficient of determination of 0.96 and a correlation coefficient of 0.98. When the same input parameters from the 15 other sampled reservoirs were input into the model, a low, moderate, or high risk of cyanotoxins was determined for each reservoir.

Hybrid Evolutionary Algorithm (HEA)

An HEA model (Grosan and Abraham 2007) uses evolutionary algorithms that apply predictive rules to limnological population dynamics. HEAs are adaptive and mimic processes of natural selection or biological evolution. They use evolutionary algorithms to recognize patterns and apply a principle of "survival of the fittest" through rules with an IF-THEN-ELSE structure when determining input parameters (Cao, et al. 2006).

(Zhang, et al. 2015)

Zhang et al. (2015) developed HEA models for 2-day-ahead cyanobacteria forecasting. An HEA model is able to represent complex relationships between multiple variables, does not require *a priori* knowledge, and uses evolutionary computations for prediction. In this paper, eight site specific, spatially-explicit models were created and one generic model using all the data was created to understand the most significant prediction variables. These models were developed using five years of data from Lake Taihu, a highly eutrophic system in China that is a water supply for millions.

The HEA is designed to improve fitness between model results and observations by evolving over time. The HEA model combines both genetic programming (GP), which is an evolutionary algorithm that optimizes model structure, and differential evolution (DE), which is an evolutionary algorithm that optimizes parameter selection. The model creates “IF,” “THEN,” “ELSE” statements of empirical equations with input parameters that are important to the prediction. Population and depth were estimated through model interaction and a sensitivity analysis was performed by changing one variable and keeping the rest constant to understand each parameter’s effect on model outputs. Additionally, a wavelet coherency analysis (the correlation between two signals at a given frequency) was performed to determine a time lag between input environmental factors and cyanobacteria response. This time lag was determined to be two days.

The spatially-explicit models of each site produced good fit to the data with an average coefficient of determination of 0.74 with a high of 0.83 and a low of 0.62 at specific sites. The generic model was tested on each site and had an average coefficient of determination of 0.57 with a high of 0.77 and a low of 0.36 at specific sites. For both types of model, phosphorus and pH were determined to be positively correlated with cyanobacteria biomass while nitrogen was determined to be negatively correlated.

(Swanepoel, et al. 2016)

Swanepoel et al. (2016) developed an HEA model to forecast future events of high cyanobacterial cell concentrations. HEAs can take highly complex ecological time-series data and successfully develop prediction rule sets. To develop the model, 10 years of physical, chemical, and biological water quality data was taken from Vaal Dam, a reservoir for drinking water in South Africa that has problems with *Microcystis*.

A principal component analysis was performed to characterize the relationships and significance of each collected variable. The HEAs then apply genetic algorithms to optimize the rule structures, using “IF,” “THEN,” “ELSE” statements and genetic algorithms for optimization of input parameters. For model training, 75% of the data were selected by bootstrapping, with the remaining 25% used for testing. The data were split so that subsets of data selected for training and testing were modeled in 50 different iterations. The best model iteration was chosen based on root mean square error and coefficient of determination.

When running the model an initial population and maximum number of repetitive runs were chosen. The HEA rule sets were discovered and then a parameter sensitivity analysis was performed by changing one variable and keeping the rest constant. The real time prediction from the model had a square correlation coefficient of 0.95 when tested with the 25% of data and a coefficient of determination of 0.97 when tested on three years of data not used in the training. The seven day in advance model produced better results than the 14 and 21 day in advance models, producing a coefficient of determination of 0.90 for the test data and 0.53 for the three years of data.

(Ostfeld, et al. 2015)

Ostfeld et al.s (2015) developed a data-driven evolutionary algorithm for cyanobacteria prediction. The model is a combination of model trees and genetic algorithms (GA) that result in a simple set of empirical linear rules. Model trees are decision trees that represent classifiers in a tree structure flow chart, and each leaf node holds a piecewise linear function. Numerical decision attributes are predicted through the linear functions and an if-then set of rules is formed. GAs are algorithms that use natural selection principles to select decision variables that are the most “fit,” or the most appropriate, for good modeling results. The model trees are used to solve classification problems, while the GAs ensure that the model is optimized.

The model was produced using physical, chemical, and biological data from Lake Kinneret in Israel. A sensitivity analysis was done by changing different input variables, the lag time, and number of generations to see which components had the highest significance on outputs and resulted in the highest correlation coefficients. Several models were created and the model of 2002-2004 had the best cross-validation correlation coefficient of 0.96.

Other

(Harris and Graham 2017)

Harris and Graham (2017) developed a paper to compare 12 unique linear and nonlinear regression modeling techniques to predict cyanobacterial abundance. Fourteen years of physiochemical water quality data from Cheney Reservoir, a reservoir used for recreation and drinking water supply, was used to build and compare models. The data were normalized and then compared using an analysis of variance, and then a data partition program was used to select a random 75% of the data to train each of the 12 predictive models. The models included, ordinary linear regression, partial least squares, elastic net, ANN, MARS, SVM, single trees, bagged trees, booted trees, conditional inference trees, random forest, and cubist models.

The SVM, random forest, and boosted trees had the lowest root mean square errors. On cyanobacterial abundances less than 60,000 cells/mL the SVM performed the best, and on concentrations greater than 60,000 cells/mL the random forest and boosted trees models were slightly better than SVM. All three models identified reservoir elevation and Chl-a as relatively important prediction variables. The cubist model (Quinlan 1993) had low performance for lower concentrations but had the overall best performance at predicting cyanobacterial abundances greater than 60,000 cells/mL. This model identified elevation and specific conductance as important prediction variables. The machine learning models were considered better cyanobacterial abundance predictors. However, no models were able to predict the highest 3% of cyanobacterial abundance.

(Mellios, et al. 2020)

Mellios and others (2020) developed and compared a stepwise multiple linear regression model and four different machine learning methods to predict cyanobacteria biomass and associated risk levels. Each model used data from 822 Northern European lakes and modeled cyanobacterial dynamics as a whole. Machine learning models were used because of their ability to learn from the data and apply associated patterns to new, unlabeled data. The model goals were to link environmental conditions to cyanobacterial concentrations and then assign an associated risk level.

Prior to modeling, a correlation matrix was produced to investigate what explanatory variables (physiochemical, meteorological, and geomorphological) were correlated with cyanobacterial biomass. Chl-a, total nitrogen, and total phosphorus were determined to be the most correlated variables. Then a stepwise multiple linear regression was performed to iteratively add and remove variables until an optimized subset was determined. The best linear model produced through this process showed low reliability in predicting cyanobacteria concentrations with a coefficient of determination of 0.44. A path analysis, which is a technique used to evaluate variable relationships, was additionally performed to decompose correlations to interpret the effects of each piece and to determine whether the variables had significant relationships with each other.

The data for each model were divided into a training (80%) and testing (20%) subset. The four machine learning models used included a Decision Tree model, a K-Nearest Neighbors model, a Support-vector Machine, and a Random Forest model. The models produced predicted values of cyanobacterial biomass. Then the predicted cyanobacterial biomass concentrations were converted to biomass cell counts so that an established high, medium, and low risk category could be applied using guidelines from the WHO (WHO 2003). Kappa statistics, which are correlation coefficients for rating determined through a confusion matrix, along with accuracy, determined by percent correct determination of high, medium, and low concentrations were used to compare models. The Random Forest model was selected as the best model, and the most optimized version of this model had accuracy as high as 95.81%.

Discussion

This review summarized 34 papers that produced mechanistic or data driven models for harmful algal bloom and cyanotoxin predictions in lakes and reservoirs. Of the 34 papers, about 70% of them were trying to predict blooms through cyanobacterial biomass and the other 30% were trying to predict cyanobacterial blooms or risk of blooms through predicted cyanotoxin concentrations. The majority of the papers (82%) developed data-driven models and no mechanistic models were found to predict cyanotoxin concentrations. This unbalance reflects a general trend of research away from mechanistic models toward machine learning approaches, in part, due to the advent of efficient machine learning algorithms and the advancement of coding, but also, it is believed, that there may be an impression that mechanistic modeling has gone as far as it can, and that HAB prediction has been a mixed success at best. Of the data-driven models, 39% of the models for cyanobacteria biomass prediction were ANN models and the SVM and MARS models were the most common models for cyanotoxin predictions (70%).

The selection process for this paper was focused on models that predicted cyanobacterial blooms in lakes and reservoirs, rather than those that focused on improving specific mechanisms (e.g., growth kinetics). Many of the mechanistic models in this literature search, derived from scientific first principles, are produced in laboratory settings for implementation in more complete transport and fate models and are not focused on bloom prediction. Thus, these papers were removed from consideration through the relevance rating process described in the methods section. Compared to the Guven and Howard (2016) review, data-driven models are becoming more frequent, likely as a consequence of the growing interest in data science approaches. The only data driven models that were reviewed in the Guven and Howard (2006) review were earlier generation ANN models in river systems.

Mechanistic models are based around differential transport, kinetic, stoichiometric, or other equations that are derived through physical, scientific theories. One of the benefits of a mechanistic model is that, because it is established on scientific principles, the model can often be applied to systems other than that for which the model was calibrated and tested, with success. They also produce output parameters, such as rate constants or dispersion coefficients, with scientific meaning in relation to the system being modeled. However, often parameters within the established equations are hard to measure or obtain independently, which can introduce uncertainties into the model itself. In many cases in this review the mechanistic models were coupled with other techniques to deal with these uncertainties. Ibelings and others (2003), for example, dealt with uncertainties produced by coupling their mechanistic model with fuzzy set theory that allowed them to establish high, moderate, and low cyanobacterial bloom risk categories.

Machine learning models often use years of observation data for training. The models learn from the input data and tune their black box equations automatically to produce optimal prediction results. The machine learning methods in this study, that rely on large amounts of data, often incorporated many measured parameters as inputs into their models. However, many of the articles showed optimization techniques used to select, up to more than 30 input parameters that would produce the best output results, while one paper, Xiao and others (2017), showed that cyanobacterial blooms could be predicted with a single input parameter. Many of the articles, especially those produced by García Nieto and others from 2011-2017, experimented with different optimization techniques to produce optimal prediction results, using the same base model.

The input parameters for the machine learning methods from different authors were very different. Some included more than 30 input parameters, some optimized their model by removing several redundant parameters, while one model only included a single input parameter. It is difficult to compare and contrast the input parameters because they are not synonymous or compatible in a way that would make them easy to compare. For example, Teles and Vasconcelos (2006) have a set of input parameters, but then split the data by years based on availability, split the input parameters to include only physical and chemical, only biological, and then all parameters, which ends up being six different models, with different input parameters. Many of the models are specific to waterbody conditions, available data, or use a wide array of optimization techniques. This makes it nearly impossible to know if the selection of data and input parameters for one machine learning method, produced using methods specific to a certain lake with certain data, is better or worse than a different model produced under a completely different set of conditions. More structure among model training methods may make models more comparable and would be beneficial to the field of cyanobacterial prediction models.

Machine learning methods are often beneficial because they are flexible and can incorporate whatever data are available into the model for prediction. They are also designed specifically for a particular system and can produce very accurate results for that system. A machine learning method may be able to produce better results for a specific lake than a mechanistic approach because of the individualistic design approach. However, being designed for one specific system is also a disadvantage as it makes the model of less value to systems other than that for which the model is trained, unlike a mechanistic model. These models also require large amounts of data to be successful and often produce results that are less expressive than mechanistic models because they are just producing a fit to observational data and may be less reliable for forecasting. Many of the papers that produced machine learning models concluded that bloom occurrence in each lake or reservoir is governed by different environmental factors and creating a universal model is nearly impossible (Ahn, et al. 2011).

Ultimately, biomass response in different water bodies can be highly variable, and each waterbody may need to have its own prediction model developed (Xu, et al. 2020). However, even though tailored machine learning models are useful for the system they are trained on during a specific time period, systems can change over time and render that model obsolete for future times under different climate and loading conditions, changes that could be captured in a mechanistic model. One option that in increasingly used in climate and weather prediction is one where a mechanistic model is used for setting up a structure for downscaling a data driven modification for a system. This hybrid approach was not specifically addressed in the articles reviewed here. However, one article (Ibelings, et al. 2003) hinted at a hybrid model combining mechanistic models with fuzzy logic to fill the gaps.

Summary and Conclusion

Cyanobacteria are organisms capable of thriving in various environmental conditions in lakes and reservoirs. They are versatile, adaptable, and can flourish under the right conditions, producing harmful algal blooms (HABs) that negatively affect aquatic life and water quality in ecosystems. These cyanobacterial blooms can deplete the waterbody of oxygen, outcompete other phytoplankton species, and produce toxins that have negative health effects on competitors and predators in the waterbody, as well as other life, including animals and humans, that use the waterbody for irrigation, recreation, or public water supply.

Decades of research to better understand cyanobacteria and their effects on aquatic ecosystems has resulted in a number of models that can accurately predict the timing of blooms and numbers or biomass of cyanobacteria, though models of cyanotoxin production have not been as well researched or successful. This paper represents a comprehensive literature review of existing models that have been developed to predict harmful algal blooms in lakes and reservoirs since about 2003. After an initial literature search seeded with keywords from previous research, 1150 articles were reviewed for inclusion in a preliminary database based on relevance to prediction of cyanobacteria or cyanotoxins under Utah lake and reservoir conditions. After several reading and relevance rating cycles, 34 articles with methods and models that were used to predict cyanobacteria and cyanotoxins were included in this review. The models discussed were developed to either predict cyanobacterial biomass or cyanotoxin concentrations in actual lakes and reservoirs. Most of the models (28) discussed in this review were data driven while only a small number (six) were mechanism-based, reflecting a change in research emphasis from prior to 2003. In addition, a majority of the model articles found were concerned with predicting cyanobacteria abundance or dominance, while only a small number were for predicting cyanotoxin production.

The data driven models require large amounts of data and are usually produced to be specific to one waterbody or relatively homogeneous group of waterbodies. The mechanistic models are developed with a scientific approach that allows them to be used for other systems with the advantage that the process of calibration may reveal the importance of the various input variables and give scientific meaning to parameter estimates. The best of these models, generally ANN or SVM, were able to explain up to 96% of the variance of cyanobacteria concentrations, though cyanotoxin prediction accuracy was significantly lower.

It is important to review past work so that future work can be optimized, and mitigation and prediction can be improved. Research must be advanced in this area so that our water resources can be protected from the harmful effects that cyanobacterial blooms create. This review updates the state of the practice of HAB modeling, based on previous work by Guven and Howard (2006), and will hopefully provide a basis for the next generation of models for mitigating the negative effects of excess HABs.

Recommendations

On the basis of this work, guidelines used to select articles should be more systematically created to filter articles earlier in the process. As it was, more than 50% of the initial set of over 1200 articles from the on-line search survived the second round of review, while the final group had only 34 articles. A clearer end goal in terms of the information incorporated into the final report would be beneficial when creating guidelines, allowing more specific factors or keywords (not only inclusive but exclusive, the second of which could have been better developed) to be identified to better sort and optimize article selection. Having a clear end goal would allow for articles that were not beneficial to be removed from the stack earlier in the process. This would help in circumventing the time-consuming aspect of reading abstracts and articles multiple times. It is also recommended that multiple search engines be used to ensure a more comprehensive aggregation of articles are found and analyzed. It would also be beneficial to include articles not written in English, as possibly beneficial articles were not included because there was no available copy of them in English.

Future work could delve into mechanistic pieces of models (e.g., kinetics, growth, toxin production, etc.) developed in laboratory settings and determine beneficial applications to lake and reservoir modeling as laboratory modeling was not included in this work. It would be beneficial for future work to focus on models that could be applicable across a wide range of waterbodies instead of overly specific to a single site. It would be useful if standard procedures or techniques were created in cyanobacterial modeling to make the process more cohesive between models and between ecosystems as well as comparable from model to model. Currently, the way the models are produced with different parameter optimization techniques, using the available data rather than the most useful data, removing or adding parameters based on arbitrary parameter limits or best exiting correlations, etc. leaves confusion about which parameters were the most beneficial to the model and if that has any scientific comparison. Another recommendation would be that future studies focus to better compare model input parameter between models and then determine how often different parameters were of interest (e.g., N, P, water temperature, etc. were good predictor parameters in a certain percentage of models produced) and determining if those parameters were different among waterbodies differing in depth, tropic status, elevation, etc.

Engineering Significance

This review was undertaken for two purposes. Foremost was to update the last comprehensive review of cyanobacterial bloom modeling efforts over the past two decades and provide a compilation that engineering practitioners can use to learn about up-to-date modeling efforts and easily find the references for those efforts. Second, it is hoped that future modeling efforts can start with this report and build on those models reported on here. This will hopefully prevent others from repeating previous work or “reinventing the wheel” and will advance the goal of predicting, mitigating, and understanding HABs.

References

ACG. (2020). “Liebig's Law of Minimum.” <https://www.acgmaterials.com/liebigs-law-minimum/> (Sep. 16, 2020).

Ahn, C. Y., Oh, H. M., and Park, Y. S. (2011). “Evaluation of environmental factors on cyanobacterial bloom in eutrophic reservoir using artificial neural networks.” *Journal of Phycology*, 47(3), 495–504.

Alonso Fernández, J. R., García Nieto, P. J., Martínez Torres, J., and Díaz Muñiz, C. (2013). “Analysis of Cyanotoxins Presence from Experimental Cyanobacteria Concentrations in the Trasona Reservoir (Northern Spain) Using Support Vector Regression.” *International Journal of Nonlinear Sciences and Numerical Simulation* 14 (2): 103–112. doi:10.1515/ijnsns-2012-0122.

Brock, T. D. (1979). *Biology of microorganisms*. Englewood Cliffs, N.J., Prentice-Hall, 1974.

Carmichael, W. W. (2001). “Health Effects of Toxin-Producing Cyanobacteria: ‘The CyanoHABs.’” Human and Ecological Risk Assessment (HERA) 7 (5): 1393–1407. doi:10.1080/20018091095087.

CDC. (2020). “General Information: Harmful Algal Bloom (HAB)-Associated Illness.” *Centers for Disease Control and Prevention* <https://www.cdc.gov/habs/general.html> (Nov. 25, 2020).

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802–813.

Elsevier. (2020). “Mendeley - Reference Management Software & Researcher Network.” Computer Software. Version 1.19.4. London, UK: Mendeley Ltd.

Elsevier. (2020). “Scopus – Abstract and Citation Database.” Computer Software.

EPA. (2009). National lakes assessment: A collaborative survey of the nation’s lakes. EPA 841-R-09-001. Washington, DC: U.S. Environmental Protection Agency, Office of Water and Office of Research and Development.

EPA. (2017). “National Water Quality Inventory: Report to Congress.” *EPA*, Environmental Protection Agency, < https://www.epa.gov/sites/production/files/2017-12/documents/305brtc_finalowow_08302017.pdf> (Oct. 27, 2020).

EPA. (2020). “Learn about Cyanobacteria and Cyanotoxins.” *EPA*, Environmental Protection Agency, <https://www.epa.gov/cyanohabs/learn-about-cyanobacteria-and-cyanotoxins> (Aug. 26, 2020).

Fadel, A., Lemaire, B. J., Vinçon-Leite, B., Atoui, A., Slim, K., and Tassin, B. (2017). “On the successful use of a simplified model to simulate the succession of toxic cyanobacteria in a hypereutrophic reservoir with a highly fluctuating water level." *Environmental Science and Pollution Research*, 24(26), 20934–20948.

García Nieto, P. J., Alonso Fernández, J. R., García-Gonzalo, E., Díaz Muñiz, C., Bayón, R. M., and González Suárez, V. M. (2015). "A New Predictive Model for the Cyanotoxin Content from Experimental Cyanobacteria Concentrations in a Reservoir Based on the ABC Optimized Support Vector Machine Approach: A Case Study in Northern Spain." *Ecological Informatics* 30: 49–59. doi:10.1016/j.ecoinf.2015.09.010.

García Nieto, P. J., Alonso Fernández, J. R., Sánchez Lasheras, F., de Cos Juez, F. J., and Díaz Muñiz, C. (2012). "A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique." *Science of the Total Environment*, 430, 88–92.

García Nieto, P. J., Fernández, J. R. A., Muñiz, C. D., Lasheras, F. S., and de Cos Juez, F. J. (2014). *Mathematical models for predicting the cyanotoxins presence in several reservoirs in the Cantabrian Basin (Northern Spain). Cyanobacteria: Ecological Importance, Biotechnological Uses and Risk Management*.

García Nieto, P. J., Fernández, J. R. A., Suárez, V. M. G., Muñiz, C. D., García-Gonzalo, E., and Bayón, R. M. (2015). "A hybrid PSO optimized SVM-based method for predicting of the cyanotoxin content from experimental cyanobacteria concentrations in the Trasona reservoir: A case study in Northern Spain." *Applied Mathematics and Computation*, 260, 170–187.

García Nieto, P. J., García-Gonzalo, E., Alonso Fernández, J. R., and Díaz Muñiz, C. (2017). "A hybrid wavelet kernel SVM-based method using artificial bee colony algorithm for predicting the cyanotoxin content from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain)." *Journal of Computational and Applied Mathematics*, 309, 587–602.

Garcia Nieto, P. J., Sánchez Lasheras, F., de Cos Juez, F. J., and Alonso Fernández, J. R. (2011). "Study of cyanotoxins presence from experimental cyanobacteria concentrations using a new data mining methodology based on multivariate adaptive regression splines in Trasona reservoir (Northern Spain)." *Journal of Hazardous Materials*, 195, 414–421.

Giannuzzi, L. (2019). "Cyanobacteria Growth Kinetics." Algae. doi:10.5772/intechopen.81545.

Greco, S., Matarazzo, B., and Slowinski, R. (2001). "Rough sets theory for multicriteria decision analysis." *European Journal of Operational Research*, 129(1), 1–47.

Grosan C., Abraham A. (2007) Hybrid Evolutionary Algorithms: Methodologies, Architectures, and Reviews. In: Abraham A., Grosan C., Ishibuchi H. (eds) Hybrid Evolutionary Algorithms. Studies in Computational Intelligence, vol 75. Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-73297-6_1

Guang-Bin, H., Qin-Yu, Z., Chee-Kheong, S. (2006). "Extreme learning machine: theory and applications". *Neurocomputing*. 70 (1): 489–501. doi:10.1016/j.neucom.2005.12.126.

Guven, B., and Howard, A. (2006). “A Review and Classification of the Existing Models of Cyanobacteria.” Progress in Physical Geography 30 (1): 1–24. doi:10.1191/0309133306pp464ra.

Harris, T. D., and Graham, J. L. (2017). “Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset.” *Lake and Reservoir Management*, 33(1), 32–48.

Hastie, T. J.; Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC. ISBN 978-0-412-34390-2.

Hayes, N. M., and Vanni, M. J. (2018). “Microcystin concentrations can be predicted with phytoplankton biomass and watershed morphology.” *Inland Waters*, 8(3), 273–283.

Ibelings, B. W., Vonk, M., Los, H. F. J., Van Der Molen, D. T., and Mooij, W. M. (2003). “Fuzzy modeling of cyanobacterial surface waterblooms: Validation with NOAA-AVHRR satellite images.” *Ecological Applications,* 13(5), 1456–1472.

Imberger J, Patterson JC (1981) A dynamic reservoir simulation model-DYRESM. In: Fischer HB (ed) Transport models for inland and coastal waters. Academic press, New York, pp 310–361.

Kennedy, J., and Eberhart, R. (1995). *Particle Swarm Optimization*. IEEE Service Center, Perth, Australia, 1942–1948.

Lee, D. T. L., and Yamamoto, A. (1994). “Wavelet Analysis: Theory and Application.” *Hewlett Packard Journal*, 45(6), 44–52.

Lou, I., Xie, Z., Ung, W. K., and Mok, K. M. (2016). “Freshwater algal bloom prediction by extreme learning machine in Macau Storage Reservoirs.” *Neural Computing and Applications*, 27(1), 19–26.

Luo, Y., Yang, K., Yu, Z., Chen, J., Xu, Y., Zhou, X., and Yang, Y. (2017). “Dynamic monitoring and prediction of Dianchi Lake cyanobacteria outbreaks in the context of rapid urbanization.” *Environmental Science and Pollution Research*, 24(6), 5335–5348.

Mellios, N., Moe, S. J., and Laspidou, C. (2020). “Machine learning approaches for predicting health risk of cyanobacterial blooms in Northern European Lakes.” *Water (Switzerland),* 12(4).

Millie, D. F., Weckman, G. R., Fahnenstiel, G. L., Carrick, H. J., Ardjmand, E., Young, W. A., Sayers, M. J., and 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.

Olsson, A. E. (2011). *Particle Swarm Optimization Theory, Techniques and Applications*. Nova Science Publishers, Hauppauge, NY.

Ostfeld, A., Tubaltzev, A., Rom, M., Kronaveter, L., Zohary, T., and Gal, G. (2015). "Coupled data-driven evolutionary algorithm for toxic cyanobacteria (Blue-Green Algae) forecasting in lake Kinneret." *Journal of Water Resources Planning and Management*, 141(4), 1–13.

Paerl, H. W., Fulton, R. S., Moisander, P. H., and Dyble, J. (2001). "Harmful Freshwater Algal Blooms, with an Emphasis on Cyanobacteria." The Scientific World Journal 1: 76–113. doi:10.1100/tsw.2001.16.

Paerl, H.W. (1982). "Chapter 3." *Advances in microbial ecology: Volume 6,* essay, Plenum Press, New York, NY, 33–64.

Phang, Y. N., and Loh, E. F. (2013). "Zero Inflated Models for Overdispersed Count Data." *International Journal of Mathematical, Computational Science and Engineering Vol:7 No:8, 2013,* 7(8), 78–80.

Quinlan, J. R. (1993). Cr.5: *Programs for machine learning*, Morgan Kaufmann, San Mateo, CA.

Randall, M. C., Carling, G. T., Dastrup, D. B., Miller, T., Nelson, S. T., Rey, K. A., Hansen, N. C., Bickmore, B. R., and Aanderud, Z. T. (2019). "Sediment potentially controls in-lake phosphorus cycling and harmful cyanobacteria in shallow, eutrophic Utah Lake." *Plos One*, 14(2).

Rhee, G.-Y. (1982). "Chapter 2. Effects of Environmental Factors and Their Interactions on Phytoplankton Growth" *Advances in microbial ecology: Volume 6,* essay, Plenum Press, New York, NY, 33–64.

Redfield, A. C, (1934). "On the proportions of organic derivatives in sea water and their relation to the composition of plankton." *James Johnstone Memorial Volume*. Liverpool University Press. 176-192

Simon, D. (2013). *Evolutionary Optimization Algorithms: biologically-inspired and population-based approaches to computer intelligence*. Wiley-Blackwell, Chichester, NY.

Specht, D. (1991). "A general regression neural network." *IEEE Transactions on Neural Networks*, 2(6), 568–576.

Steffen, M. M. (2017). "Ecophysiological Examination of the Lake Erie Microcystis Bloom in 2014: Linkages between Biology and the Water Supply Shutdown of Toledo, OH." *Environmental Science & Technology*, 51(12), 6745–6755.

Swanepoel, A., Barnard, S., Recknagel, F., and Cao, H. (2016). "Evaluation of models generated via hybrid evolutionary algorithms for the prediction of Microcystis concentrations in the Vaal Dam, South Africa." *Water SA*, 42(2), 243–252.

Tao, M., Xie, P., Chen, J., Qin, B., Zhang, D., Niu, Y., Zhang, M., Wang, Q., and Wu, L. (2012). "Use of a generalized additive model to investigate key abiotic factors affecting microcystin cellular quotas in heavy bloom areas of lake Taihu." *PLoS ONE*, 7(2).

Taranu, Z. E., Gregory-Eaves, I., Steele, R. J., Beaulieu, M., and Legendre, P. (2017). “Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor.” *Global Ecology and Biogeography*, 26(6), 625–637.

Teles, L. O., Vasconcelos, V. (2006). “Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks.” *Environmental Management,* 38(2)*,* 227–237.

UDEQ. (2018). “Rules and Regulations: Drinking Water HABs Response Plan.” <https://deq.utah.gov/drinking-water/rules-and-regulations-habs> (Aug. 25, 2020).

UDEQ. (2019). “Utah Lake Algal Bloom Monitoring 2019.” *Utah Department of Environmental Quality*, <https://deq.utah.gov/water-quality/utah-lake-algal-bloom-monitoring-2019> (Sep. 2, 2020).

UDEQ. (2020). “2020 Recreational Health Advisory Guidance for Harmful Algal Blooms.” *Utah Department of Environmental Quality*, <https://deq.utah.gov/water-quality/recreational-health-advisory-guidance> (Aug. 25, 2020).

UDEQ. (2020). “Division of Drinking Water Harmful Algal Bloom & Cyanotoxin Response Plan.” *Utah Department of Environmental Quality*, <https://deq.utah.gov/drinking-water/division-drinking-water-harmful-algal-bloom-cyanotoxin-response-plan> (Aug. 26, 2020).

Vapnik, V. N. (1995). “The Nature of Statistical Learning Theory.” *Springer*. New York, NY, USA.

Vinçon-Leite, B., Fadel, A., Lemaire, B. J., Bonhomme, C., Li, Y., Le Divechen, G., Zhang, J., and Luo, Y. (2017). “Short-term forecasting of cyanobacteria blooms in Yuqiao reservoir, China.” *Houille Blanche*, 2017-April(2), 35–41.

Von Altrock, C. (1995). “Fuzzy logic and neuro fuzzy applications explained. Prentice Hall, Upper Saddle River, New Jersey, USA.

Wang, H., Zhizhang, Z., Zhao, Y., and Dongfang, L. (2016). “Projection pursuit-based Microcystis bloom warning in a Riverside Lake.” *Water, Air, and Soil Pollution,* 227(4).

Wang, L., Kang, J., Xu, J., Zhang, H., Wang, X., Yu, J., Sun, Q., and Zao, Z. (2020). “Early warning of cyanobacteria blooms outbreak based on stoichiometric analysis and catastrophe theory model.” *Journal of Mathematical Chemistry,* 58(5), 906–921.

Wang, Z., Huang, K., Zhou, P., and Guo, H. (2010). “A hybrid neural network model for cyanobacteria bloom in Dianchi Lake.” *Procedia Environmental Sciences,* 2(5), 67–75.

Wang, Z., Li, Z., and Li, D. (2012). “A niche model to predict Microcystis bloom decline in Chaohu Lake, China.” *Chinese Journal of Oceanology and Limnology,* 30(4), 587–594.

World Health Organization (WHO). (2003). “Guidelines for Safe Recreational Waters: Coastal and Fresh Waters.” Chapter 8. *Coastal and Fresh Waters*, World Health Organization, Geneva, Switzerland, 136–158.

Xiao, X., He, J., Huang, H., Miller, T. R., Christakos, G., Reichwaldt, E. S., Ghadouani, A., Lin, S., Xu, X., and Shi, J. (2017). “A novel single-parameter approach for forecasting algal blooms.” *Water Research*, 108, 222–231.

Xie, Z., Lou, I., Ung, W. K., and Mok, K. M. (2012). “Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs.” *Mathematical Problems in Engineering*, 2012.

Xu, T., Yang, T., and Xiong, M. (2020). “Time scales of external loading and spatial heterogeneity in nutrients-chlorophyll a response: Implication on eutrophication control in a large shallow lake.” *Ecological Engineering*, 142.

Yang, X. S., and Deb, S. (2010). “Engineering optimisation by cuckoo search.” *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(4), 330–343.

Zeeman, E. C. (1976). “Catastrophe Theory.” *Scientific American*, 66–70-75–83.

Zhang, X., Recknagel, F., Chen, Q., Cao, H., and Li, R. (2015). “Spatially-explicit modelling and forecasting of cyanobacteria growth in Lake Taihu by evolutionary computation.” *Ecological Modelling*, 306, 216–225.

Zhao, C. S., Shao, N. F., Yang, S. T., Ren, H., Ge, Y. R., Feng, P., Dong, B. E., and Zhao, Y. (2019). “Predicting cyanobacteria bloom occurrence in lakes and reservoirs before blooms occur.” *Science of the Total Environment*, 670, 837–848.

Zhao, C. S., Yang, S. T., Liu, C. M., Dou, T. W., Yang, Z. L., Yang, Z. Y., Liu, X. L., Xiang, H., Nie, S. Y., Zhang, J. L., Mitrovic, S. M., Yu, Q., and Lim, R. P. (2015). “Linking hydrologic, physical and chemical habitat environments for the potential assessment of fish community rehabilitation in a developing city.” *Journal of Hydrology*, 523, 384–397.

Zhao, C., Sun, C., Liu, C., Xia, J., Yang, G., Liu, X., Zhang, D., Dong, B., and Sobkowiak, L. (2014). “Analysis of regional zoobenthos status in the Huai River Basin, China, using two new ecological niche clustering approaches.” *Ecohydrology*, 7(1), 91–101.

Zuur, A. F., and Ieno, E. N. (2016). *Beginner's guide to zero-inflated models with R*. Highland Statistics Ltd., Newburgh, United Kingdom.

Appendices

Appendix A – Important Key Words

Important Keywords			
Cyanotoxins	Growth	Modelling	Freshwater
Cyanobacteria	Growth Limitation	Modeling	Toxins
Cyanobacterium	Growth Relationship	Mathematical Models	Toxicity
Cyanobacterial Bloom	Growth Rates	Monitoring	Toxin Production
Bloom	Growth Kinetics	Lakes	Harmful Algal Bloom
Predicting Blooms	Growth Inhibition	Reservoirs	Microalgae
Causes of Bloom	Algae	Phytoplankton	Green algae
Epilithic Cyanobacteria	Algal Bloom	Zooplankton	Blue Green Algae
Benthic Cyanobacteria	Algae Blooms	Photosynthesis	Cylindrospermopsin
Cyanobacterial Growth	Algal Concentration	Water Quality	Anabaena
Cyanobacterium Growth	Nuisance Algae	Water Quality Management	Aphanizomenon
Toxic Benthic Cyanobacteria	Harmful Algae	Microcystins	Dolichospermum
Cyanobacteria Monitoring	Epiphytic Algae	Microcystis	Heterocysts
Limnology	Nutrients	Water Temperature	Nodularin
CyanoHABs	Nitrogen	Temperature	Prediction
Cyanophytes	Nitrates	Chlorophyll	Predictive Power
Eutrophication	Phosphorus	Forecasting	Predictive Success
Freshwater Eutrophication	Model	Lake Eutrophication	

Appendix B – Article Database and Selected Articles

The following link is associated with a google sheet with the entire article data base (Sheet 1 – All Articles) and the 34 articles found summarized in this report (Sheet 2 – Selected Articles).

https://docs.google.com/spreadsheets/d/1fuHx9smoJsb7Bj12FFabsV9yROlmt8i_yJaaYUCNLao/edit#gid=0

Appendix C – Acronyms and Abbreviations

ABC	Artificial Bee Colony
AGM	Adaptive Grey Model
AICc	Corrected Akaike Information Criteria
ANN	Artificial Neural Network
BIC	Bayesian Information Criterion
BP-ANN	Back-Propagation Neural Network
BRT	Boosted Regression Tree
CCA	Canonical Correspondence Analysis
Chl-a	Chlorophyll a
DE	Differential Evolution
DEQ	Department of Environmental Quality
DIC	Dissolved Inorganic Carbon
DYRESM-	Dynamic Reservoir Simulation Model-
CAEDYM	Computational Aquatic Ecosystem Dynamics Model
ELM	Extreme Learning Machine
GA	Genetic Algorithm
GAM	Generalized Additive Model
GLMM	Generalized Linear Mired Model
GP	Genetic Programming
GRNN	Generalized Regression Neural Network
HABs	Harmful Algal Blooms
HALs	Health Advisory Levels
HEA	Hybrid Evolutionary Algorithm
MARS	Multivariate Adaptive Regression Spline
MLP	Multilayer Perceptron
NLA	National Lake Assessment
PSO	Particle Swarm Optimization
RD	Rough Decision
SOM	Self-Organizing Map
SVM	Support Vector Machine
SVR	Support Vector Regression
TI-TP	Diatom-Inferred Total Phosphorus
UDEQ	Utah Department of Environmental Quality
WA:SA	Watershed Area to Lake Surface Area
WHO	World Health Organization
WNN	Wavelet Analysis Artificial Neural Network