Leveraging Hydroclimatic Processes and Remote Sensing for Biological Response in Water Resource Management: Applications to Water Quality and Water-related Disease

by
Maxwell R.W. Beal

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Civil and Environmental Engineering)

At the
UNIVERSITY OF WISCONSIN – MADISON

2024

Date of final oral examination: 04/30/2024

The dissertation is approved by the following members of the Final Oral Committee:
Paul Block, Associate Professor, Civil and Environmental Engineering
Katherine McMahon, Professor, Civil and Environmental Engineering
Mutlu Özdoğan, Associate Professor, Forest and Wildlife Ecology
Ajay Sethi, Professor, Population Health Sciences
Grace Wilkinson, Associate Professor, Integrative Biology

Dedication

To the managers, decision-makers, and stewards of our water resources.

Acknowledgements

My years pursuing a PhD at UW-Madison have been incredibly rewarding, challenging, and fulfilling. My experiences with the faculty and students of UW – Madison have shaped my interests, developed my scientific identity, and provided endless opportunities for personal growth. This dissertation would not be possible without the continual support I received from mentors, collaborators, peers, and friends. To my advisor, Dr. Paul Block, thank you for a wonderful graduate experience. I am so grateful for your guidance over the past five years. Your commitment and enthusiasm for teaching, research, and mentorship have given my interest in science a purpose and direction, and I am eager to pursue the example you set. To my committee members Dr. Trina McMahon, Dr. Mutlu Özdoğan, Dr. Grace Wilkinson, and Dr. Ajay Sethi, thank you for your guidance. I have enjoyed collaborating and learning from each of you throughout my time at UW-Madison. Thank you all for making me a better scientist, your contributions and feedback have been crucial to this work.

This work is supported by an NSF CAREER project grant (1845783) and the UW-Madison graduate school. Thank you to the North Temperate Lakes Long Term Ecological Research Program for providing much of the water quality data used in this work and support, and to the following field technicians for their contributions to the project: Lauren Belz, Maddie Douglas, Lucy Pine, and Erin Rice. I also acknowledge the Global Health Institute and the One Health Center – Colombia for their contributions and support of my final chapter.

Thank you to everyone in the Water Systems and Society lab and the Water Resources Engineering program for your friendship and support. To my friends and family, I could not have done it without you. I am so grateful to have shared this time with many of you in Madison.

Table of Contents

List of Figures

List of Tables

Abstract

Hydrology and climate strongly influence the distribution and abundance of living organisms. Hydroclimate variability on annual and seasonal timescales, and the occurrence of hydroclimate extremes, structure many biological systems, particularly those closely related to water. The indirect impacts of hydroclimate variability on biological responses, while complicated, can have pronounced impacts on public health. Two such responses include water quality and water related disease. In many regions variability in water quality and water related disease presents a notable challenge to water managers and public health officials who must allocate limited resources to manage uncertain outcomes. To adequately address the consequences of uncertainty in water availability (quality and quantity) on ecosystem and public health, officials are increasing demanding novel methods to manage water, and downstream impacts, in vulnerable regions. This dissertation explores the application of hydroclimatic processes and satellite remote sensing for the development of comprehensive sub-seasonal to seasonal forecasts and satellite-based monitoring systems to improve management, resource allocation, and public health related to lake water quality and water related disease. Four chapters are presented: Chapter one explores the development and assessment of a sub-seasonal and seasonal cyanobacteria forecasting system conditioned on local, global, and within-lake predictors for Lake Mendota, a small eutrophic lake in Madison, Wisconsin. Chapter two identifies relevant hydroclimate predictors and develops season-ahead forecasts for peak season harmful algae metrics in 178 lakes across the Northeast and Midwest U.S., illustrating the potential for implementation of skillful seasonal water quality forecasts at scale. Chapter three investigates the potential for satellite-based monitoring of harmful algae indicators on Lake Mendota, and chapter four pivots to understand how hydroclimate variables influence water-related disease in the Amazon basin, and how season-ahead forecasts of disease can be used to inform public health interventions.

Chapter 1. Introduction

Hydroclimatic variability has a significant impact on the function and management of many natural and human systems. In particular, increases in the intensity and frequency of hydroclimatic extremes related to climate change are likely to inflate hydroclimate-driven risks to human and ecosystem health (IPCC, 2012; Stevenson et al., 2022). In the United States, 2021 produced 20 individual billion-dollar weather and climate events including severe storms, tropical cyclones, wildfires, droughts, and floods (NCEI, 2022). Beyond the immediate risks associated with such extremes, downstream effects are shown to influence extreme behavior in variables such as water quality (Carpenter et al., 2015, 2018; Sinha et al., 2017; Towler et al., 2010), and water-related disease (Craig et al., 1999; Sena et al., 2015; Teklehaimanot et al., 2004). Thus, understanding hydroclimatic variability is critical for the successful management of a wide array of hydroclimate-driven risk factors. Advanced notice of extreme conditions can be highly advantageous for allocating resources to mitigate harm. However, for variables such as water quality and water-related disease, gaps exist in the understanding of hydroclimatic influence, application of hydroclimate variables for prediction, and evaluation of predictions coupled with decision-models to prescribe early actions.

For many waterbodies, hydroclimatic variability plays an important role in determining water quality on inter- and intra-annual timescales, and may influence the suitability of conditions for algae growth (León-Muñoz et al., 2018; Scordo et al., 2022). Harmful algae has become a prominent water quality concern in recent years as anthropogenic disturbance of nitrogen and phosphorus cycles has resulted in widespread eutrophication, leading to an increase in the prevalence of algae in many waterbodies (O’Neil et al., 2012; Paerl & Paul, 2012; V. H. Smith, 2003). In particular, the proliferation of algae in surface freshwaters has negative consequences for ecosystem function (Huisman et al., 2018; Sunda et al., 2006), economic opportunity (Dodds et al., 2009), and human health due to the potential for toxin production in some species, including cyanobacteria (Carmichael, 2001; Carmichael & Boyer, 2016). A wide array of physical, chemical, and biological processes influence algae biomass, however, the environmental signals from hydrology and climate information are often more stable than within lake processes over long time scales, potentially providing managers with important information on water quality conditions at seasonal lead times. Considerable progress has been made in the development of season-ahead forecasts to address water quantity management (e.g., Chiew et al., 2003; Delorit et al., 2017; Baker et al., 2019; Giuliani et al., 2019), however significantly less focus has been devoted to the application of season-ahead forecasts for water quality management. Longer-lead (months) pre-season predictions of expected algae conditions may allow lake managers to address a different set of actions (e.g. life-guard training, public awareness, etc.) and decisions, (e.g., testing and monitoring budgets and plans) than short term predictions.

Seasonal forecasts are also applicable to other water-related public health threats, including the spread of water-related disease. In particular, the use of hydroclimatic predictors to develop malaria early warning systems has received significant focus (L. R. Beck et al., 1994; Craig et al., 1999; Midekisa et al., 2012; Sena et al., 2015; Wimberly et al., 2022). Seasonal forecasts provide management information at a timescale that is effective for budgeting and allocation of resources, however, the use of malaria early warning systems to implement optimal early action thresholds for disease reduction has received relatively less attention. Exploration of forecast based early action with regard to water-related disease may reduce morbidity, mortality, and reduce costs for relief organizations, and may help promote forecast uptake by management agencies.

Combining seasonal forecasts with short term (day to weeks) monitoring strategies provides managers with a suite of tools that can be used to take actions at several timescales. Due to the cost-effectiveness and temporal consistency of satellite image data, remote sensing has become a powerful tool for water quality monitoring in particular (Yan et al., 2018). In remote sensing, chlorophyll-*a* and phycocyanin are often used as surrogates for cyanobacteria abundance, chlorophyll-*a* representing all phytoplankton and phycocyanin being characteristic of cyanobacteria (Dekker, 1993). Both chlorophyll-*a* and phycocyanin may be useful from a management perspective. Stumpf et al., (2016), found both chlorophyll-*a* and phycocyanin to be useful in retrieval of cyanobacterial toxins. Additionally, the ability to discriminate between cyanobacteria and other algae species during a bloom event may allow water managers to make informed decisions about closing beaches or communicating water quality information to the public.

In this dissertation, I broadly address the theme of *how hydroclimate and remotely sensed information can be integrated into decision-making tools to better anticipate and manage extreme biological outcomes in water resources systems*. Development of novel tools to address biological outcomes occurring downstream of hydroclimate variability is an important piece of managing public and ecosystem health under increasing climate uncertainty. The specific objectives of this research theme are to:

1) Develop and assess targeted season-ahead forecasts of biological outcomes conditioned on hydroclimate variables for use in management of water quality and water-related disease.
2) Explore the ability of satellite remote sensing to retrieve water quality metrics and differentiate between harmful algae indicators to improve monitoring capabilities.

These objectives are evaluated by addressing the following set of questions:

- Can hydroclimate information be used to skillfully predict harmful algae outcomes and beach closings at seasonal lead times in small inland lakes? (Chapter 2)
- Are hydroclimate-based seasonal harmful algae forecasts transferable to other lakes at scale, and what lake characteristics are associated with skillful models (Chapter 3)?
- Can satellite imagery and machine learning methods improve monitoring and discrimination of algae on a small inland lake (Chapter 4)?
- Can hydroclimate information be used to skillfully predict dengue virus incidence, and if so, in what scenarios is hydroclimate information useful (Chapter 5)?

Addressing these research questions also provides insight into relationships between global atmospheric and oceanic systems, local hydroclimatic conditions, and ecosystem function as they relate to biological outcomes of concern. Research objectives are largely pursued using two case studies: Lake Mendota, in Madison, WI, USA and four cities across Colombia including Cali, Medellin, Cucuta, and Leticia. These locations are both data rich and are the focus of active public health management efforts. I introduce these case studies and further motivation in the following two sections.

1.1 Study Background: Lake Mendota, WI

Located in Madison, Wisconsin, Lake Mendota's 596-square kilometer watershed is highly urbanized (21%) and agricultural (53%) (Betz & Genskow, 2012) (Figure 1). Municipal wastewater discharge fueled eutrophication in Lake Mendota from the 1940's-1970's, however, in recent years urban and agricultural development in the Mendota watershed has maintained the state of high productivity in the lake (R C Lathrop et al., 1998; Richard C Lathrop, 2007). In recent decades, cyanobacteria blooms have become a common summertime phenomenon (Brock, 1985; Lathrop and Carpenter, 1992, Lathrop et al., 1998). In Lake Mendota, cyanobacteria biomass is generally most apparent from June-August (Figure 2), with elevated levels typically observable in July and August.

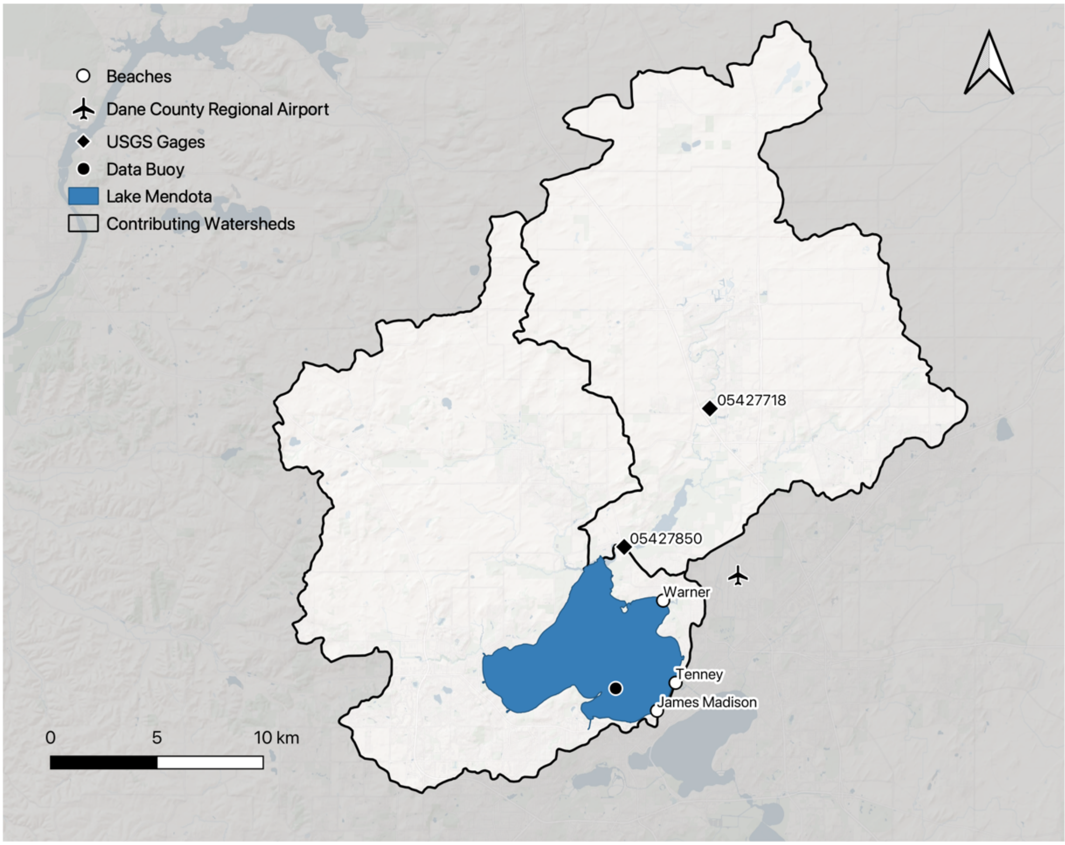

Figure 1-1. Lake Mendota and the Mendota watershed with select beaches, USGS gages and the North Temperate Lakes Long Term Ecological Research (NTL-LTER) data buoy indicated (Basemap: Carto, Watershed: WI DNR 2020, Lake: City of Madison 2019)

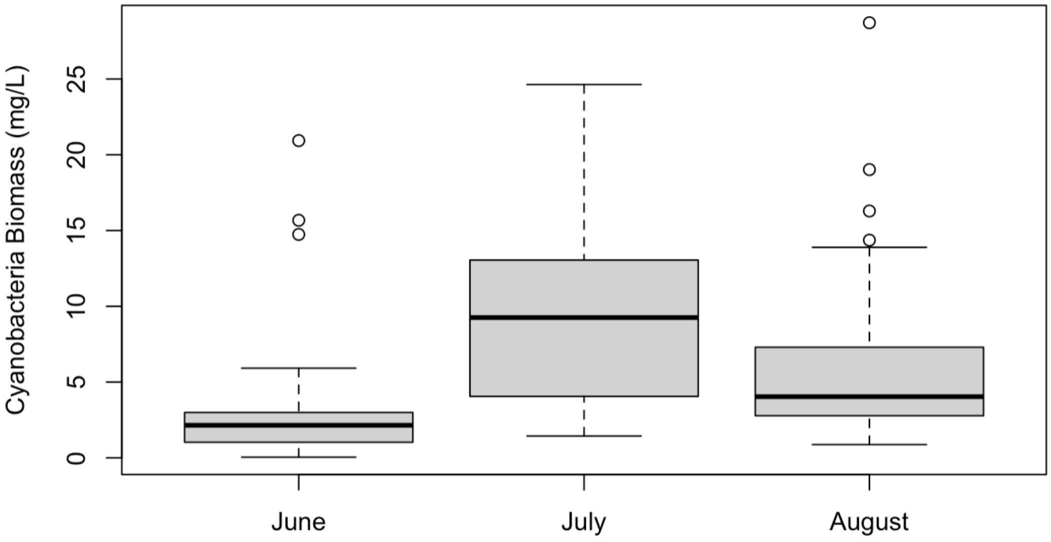

Figure 1-2. June, July, and August average cyanobacteria biomass at the LTER buoy in Lake Mendota (1995-2017). Aggregated from composite samples 0 – 8 meters of depth.

Numerous large-scale climate phenomena influence climate conditions in the upper Midwestern US, however one of the most prominent teleconnection patterns affecting precipitation and temperature is the El Niño Southern Oscillation (ENSO) (Ropelewski & Halpert, 1987). ENSO's influence globally is widely studied and generally understood, acting independently or interacting with other large-scale climate phenomena such as the Pacific Decadal Oscillation (Kahya & Dracup, 1993; Shabbar & Skinner, 2004). The state of these climate phenomena and the atmospheric-oceanic system is an important factor in local climate variability, and therefore regulate many processes important to cyanobacteria growth (Justić et al., 2005; M. Zhang et al., 2012).

Lake Mendota is an ideal candidate for the development and assessment of season-ahead forecasts and remote sensing methods for water quality monitoring. Mendota is often labeled as one of the most studied lakes in the world (Aoki, 1989; Brezonik & Lee, 1968; Konopka & Brock, 1978;

Lathrop, 2007) and is included in the National Science Foundation's North Temperate Lakes - Long-Term Ecological Research (LTER) network and has a wealth of high-quality, long-term ecological data (Magnuson et al., 2022). Rich water quality datasets allow for robust evaluation of the connectedness between hydroclimatic processes and cyanobacteria growth, and the efficacy of satellite remote sensing for retrieval of harmful algae indicators.

1.2 Study Background: Colombia

Four cities in Colombia are chosen as study sites for the development of season-ahead dengue forecasts: Cali, Medellin, Cucuta, and Leticia (Figure 3). Dengue has spread significantly since its re-emergence in Latin America, with cases rising rapidly since the 1980's (Lenharo, 2023). Colombia in particular is experiencing a resurgence of vector-borne diseases, and has been identified as an emerging disease hotspot (Jones et al., 2008). Colombia has recognized dengue virus as a significant public health threat since the 1950's. The suspension of vector control campaigns targeting *Aedes* mosquitos in 1970 led to a resurgence of dengue infections that persists today (Gutierrez-Barbosa et al., 2020). The majority of cases come from the urban areas of Colombia, driven in part by high population density and water infrastructure that may act as breeding sites for *Aedes aegypti* (Villar et al., 2015). Dengue virus is considered hyperendemic in Colombia, due to the co-circulation of all four dengue virus serotypes.

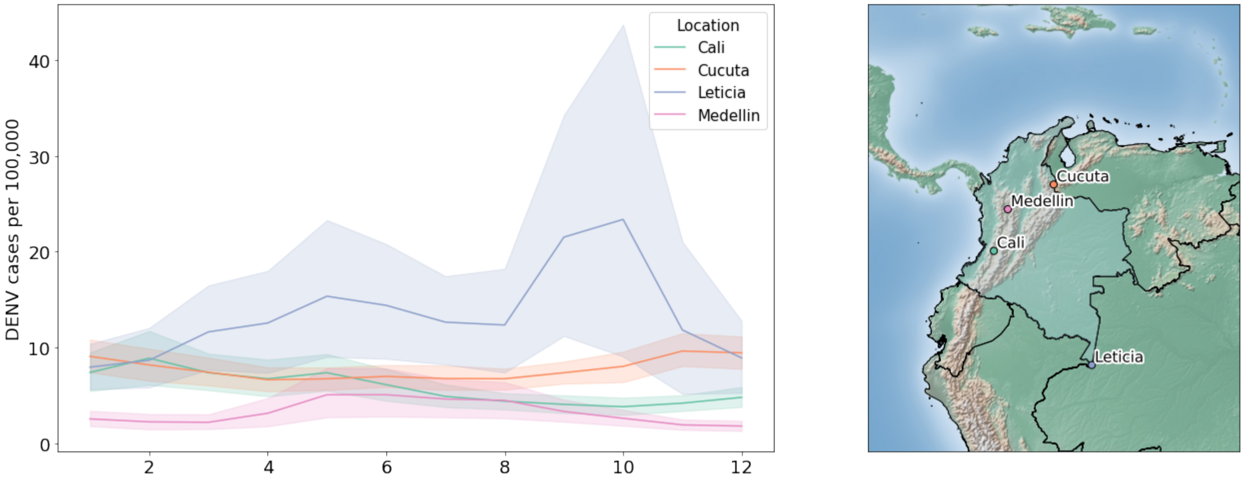

Figure 1-3. Average monthly dengue virus incidence per 100,000 population (i.e., climatology, left) and study site locations (right).

Significant portions of the country are favorable to transmission of vector-borne disease (Cabrera & Selvaraj, 2020) and variability in topography, hydrology, and climate across Colombia contributes to significant temporal and spatial variability in dengue transmission. A long-running dengue surveillance effort exists in Colombia, run by the National Public Health Surveillance System (SIVIGILA), under the National Institutes of Health of Colombia (INS). Case counts are reported by clinics and hospitals to insurance agencies and regional health authorities, which are then sent to INS for consolidation. These factors make the region a suitable study site for the investigation of hydroclimate-vector relationships and the development of dengue forecasts. Season – ahead forecasts aim to provide public health officials with information at longer lead times to aid in decision-making regarding activating public health interventions. While a dengue vaccine (Dengvaxia) is newly available in the region, traditional dengue prevention measures, including use of larvicide, insecticide, and personal mosquito bite prevention measures, remain a viable and cost-effective option to reduce dengue incidence (Claypool et al., 2021; Ocampo et al., 2014; Sepulveda & Vasilieva, 2016). These interventions may benefit from advanced notice of increased dengue risk at seasonal leads.

Chapter 2. Development of a seasonal to sub-seasonal cyanobacteria biomass forecast system for Lake Mendota, WI

Part A. A Season-ahead Cyanobacteria and Beach Closing Forecast

Adapted from: Beal, M. R. W., O'Reilly, B. E., Soley, C. K., Hietpas, K. R., & Block, P. J. (2022). Variability of summer cyanobacteria abundance: can season-ahead forecasts improve beach management? Lake and Reservoir Management, 1-16. DOI: 10.1080/10402381.2022.2084799

2.A.1 Introduction

Cyanobacteria represent some of the most ancient microorganisms on Earth, having appeared roughly 2.7 billion years ago (Schirrmeister et al., 2013). In recent decades, accelerated nutrient input and widespread land cover change have resulted in a rapid expansion of harmful cyanobacteria in our coastal waters and inland lakes (Taranu et al., 2015). Cyanobacteria are photosynthetic bacteria that thrive in eutrophic waterbodies characterized by large influxes of nutrients. Cyanobacteria can form mats known as harmful algal blooms (HABs) triggering concern from health officials and water managers given their widely identified negative ecological, aesthetic, and socioeconomic implications (Dodds et al., 2009; Huisman et al., 2018; Paerl, 2017). Importantly, common species of cyanobacteria (e.g. *Microcystis*) produce hepato- and neurotoxins, threatening waterbodies used for recreation and drinking water (Taranu et al., 2012; Timothy T Wynne & Stumpf, 2015). The negative impacts of HABs have received notable attention in larger waterbodies, such as Lake Erie. In 2014 Toledo, Ohio was forced to issue a 'do not drink' advisory due to dangerous concentrations of cyanobacteria produced toxins in the public water supply (Bullerjahn et al., 2016). Widespread eutrophication, climate change, and an established relationship between algal biomass and nutrient input suggest that cyanobacteria pose a significant threat to small inland lakes, which have so far received less attention (Huisman et al., 2018; Paerl & Huisman, 2008; V. H. Smith, 2003).

Many meteorological, chemical, and biological variables influence cyanobacteria abundance, creating a complex, dynamic ecosystem (Hamilton et al., 2016; C A Stow et al., 1997). Additionally, each cyanobacteria community has its own unique characteristics, further complicating the understanding of how concurrent drivers collectively impact cyanobacteria population in a lake (P A Soranno et al., 1997; Taranu et al., 2012). For species commonly found in eutrophic dimictic lakes, key factors influencing overall abundance include temperature, water column stability, wind, nutrient availability, precipitation, atmospheric pressure, light transparency, and predatory grazing (Ostfeld et al., 2015; Paerl, 1988; P A Soranno et al., 1997; Taranu et al., 2012). Cyanobacteria abundance expresses notable intra-seasonal variability, typically peaking during the summer season, and inter-annual variability (Richard C Lathrop & Carpenter, 1992). Lake and beach managers, however, often have limited access to information indicating the expected intensity of cyanobacteria abundance ahead of the peak season for cyanobacteria productivity. Reactive management operations, in such cases, are used in determining recreational safety and beach closures. Advanced notice of increased cyanobacteria abundance may allow lake and beach managers to alter cyanotoxin testing routines, train and inform lifeguards to watch for dangerous algae conditions, and launch public awareness campaigns before the high-risk season. Seasonal forecasts are intended to work in concert with shorter-term forecasts (days to weeks), providing managers with information at several timescales. Providing decision-makers with information on cyanobacteria conditions ahead of the high-risk season aims to extend an existing system of forecasts to improve overall management of cyanobacteria.

In recent decades, season-ahead forecasts have become a focus of research in many fields, with significant effort put towards predicting average or extreme precipitation, discharge, and temperature to inform operations in agriculture and reservoir management (Hansen et al., 2011; Wood et al., 2005). Forecasts at this scale typically aim to provide information characterizing the upcoming season, not a prescription of when events will occur. Many important management decisions fall into the gap between short-term and long-term forecasts. The development of forecasting systems at monthly and seasonal timescales can strengthen disaster preparedness by informing long-term contingency plans and activating short-term early warning systems (Vitart et al., 2012). In contrast to water quantity, relatively little attention has been devoted to season-ahead prediction of water quality.

Currently, several short-term cyanobacteria forecasts are available through entities such as the National Oceanic and Atmospheric Administration (NOAA) for the purpose of beach management (Kavanaugh et al., 2013). Forecasts are issued up to 5 days out, based on local meteorological conditions and high-resolution satellite imagery. A review of forecast and predictive models for cyanobacteria blooms found that most existing models operate on forecast horizons of less than one week, with very few extending beyond 30 days (Rousso et al., 2020). Existing season-ahead forecasts of cyanobacteria abundance have been developed with a focus on spring phosphorus loads (e.g., by NOAA Great Lakes Environmental Research Laboratory) primarily to determine necessary nutrient reductions for targeted local management plans (Obenour et al., 2014; C A Stow et al., 1997). Phosphorus is generally accepted as the limiting nutrient for cyanobacteria growth in freshwater systems and has received significant attention in seasonal forecasting due to the importance of phosphorus management in many watersheds (Downing et al., 2001; R C Lathrop et al., 1998; David W Schindler, 1977; V. H. Smith, 2003). The abundance of cyanobacteria, however, is controlled by the dynamic state and reactions of many physical, chemical, and biological variables during both the prior and concurrent seasons creating a complex array of ecosystem processes (Ostfeld et al., 2015; B. Zhu et al., 2019). Phosphorus is widely accepted as a driver of cyanobacteria productivity, and strong correlations between phosphorus load and cyanobacteria biomass have been demonstrated (R C Lathrop et al., 1998; V. H. Smith, 1985; C A Stow et al., 1997). Local hydroclimatic processes, such as extreme rain events and river discharge, may influence phosphorus loading during the spring (Carpenter et al., 2018). Spring and summer temperatures may also control cyanobacteria productivity through direct effects on photosynthetic capacity, influencing competition with other photosynthetic organisms (Taranu et al., 2012). Therefore, consideration of season-ahead, local and large-scale hydroclimatic drivers may have potential to improve the skill of season-ahead cyanobacteria forecasts.

The application of season-ahead forecasts to beach management allows for the investigation of season-ahead, non-anthropogenic drivers of cyanobacteria abundance. Incorporation of hydroclimatic (e.g., non-manageable) variables may allow for skillful forecasts of cyanobacteria abundance at seasonal timescales. The focus of this chapter is to build and assess the skill of season-ahead cyanobacteria abundance forecasts conditioned on local and global scale hydroclimatic predictors, and the subsequent ability of seasonal forecasts to predict beach closings.

2.A.1.2 Study Site

With the University of Wisconsin-Madison on its southern shore, Lake Mendota in Madison, Wisconsin (Figure 1-1) is one of the most studied ecosystems on the planet (Stephen R Carpenter et al., 2006; Richard C Lathrop, 2007). The lake covers roughly 40 km^2 and is the first of four lakes in the Yahara River basin. The 596 km^2 Mendota watershed is 21% urban and 53% agricultural (Betz & Genskow, 2012). Mendota has a long history of eutrophication dating back to the 1940s, although anecdotal evidence of cyanobacteria blooms can be found as early as the 1880s (Richard C Lathrop, 2007). From the 1940s until the 1970s, high nutrient concentrations were fueled by municipal wastewater discharge, however, Lake Mendota remains highly eutrophic to this day due to agricultural and urban development (R C Lathrop et al., 1998).

Today, most nutrient conveyance is the result of manure application in the upper part of the Yahara watershed (Betz & Genskow, 2012). Nutrient concentrations have been the focus of multiple cyanobacteria prediction models on Lake Mendota. An existing prediction model, developed by Stow et al. (1997) applied spring center-of-lake phosphorus to predict summertime cyanobacteria biovolume, with some success. Additionally, Lake Mendota was included in a Bayesian network model developed to assess the influence of short-term (1-2 weeks) nutrient concentrations (nitrogen and phosphorus) and climatic variables (air temperature, sunlight, and wind speed) on the probability of cyanobacteria blooms (Rigosi et al., 2015). To build on these efforts, the forecast presented here investigates the influence of local and global hydrologic and climatic drivers of cyanobacteria biomass at a seasonal timescale and creates a tool for proactive lake and beach management. This forecast works in concert with a sub-seasonal forecasting model for July-August cyanobacteria abundance, developed by Beal et al. (2021). Cyanobacteria prediction modeling at both the seasonal and sub-seasonal time scale allows for the consideration of both pre-season (March-May) and within-season (June) drivers of productivity and provides lake managers with two opportunities to adjust management strategies before cyanobacteria abundances peak for the summer. As a cornerstone of the Madison community, millions of dollars have been invested in water quality monitoring offering a uniquely rich data set. Mendota is therefore well-suited as a test case for the development, evaluation, and implementation of a season-ahead cyanobacteria forecasting system. Water quality data is available through the Northern Temperate Lakes – Long Term Ecologic Research (NTL-LTER) database (Magnuson et al., 2022), and beach-closing data is available by request from the Madison-Dane County Public Health department (PHMDC, 2020). Cyanobacteria abundance and associated beach closings typically peak across the June – August (JJA) summer season, with the greatest abundances typically occurring between July and August. The forecast developed here addresses average summertime (June-August) cyanobacteria biomass to inform lake and beach management decisions at the beginning of the peak season for cyanobacteria productivity (Figure 2A-1).

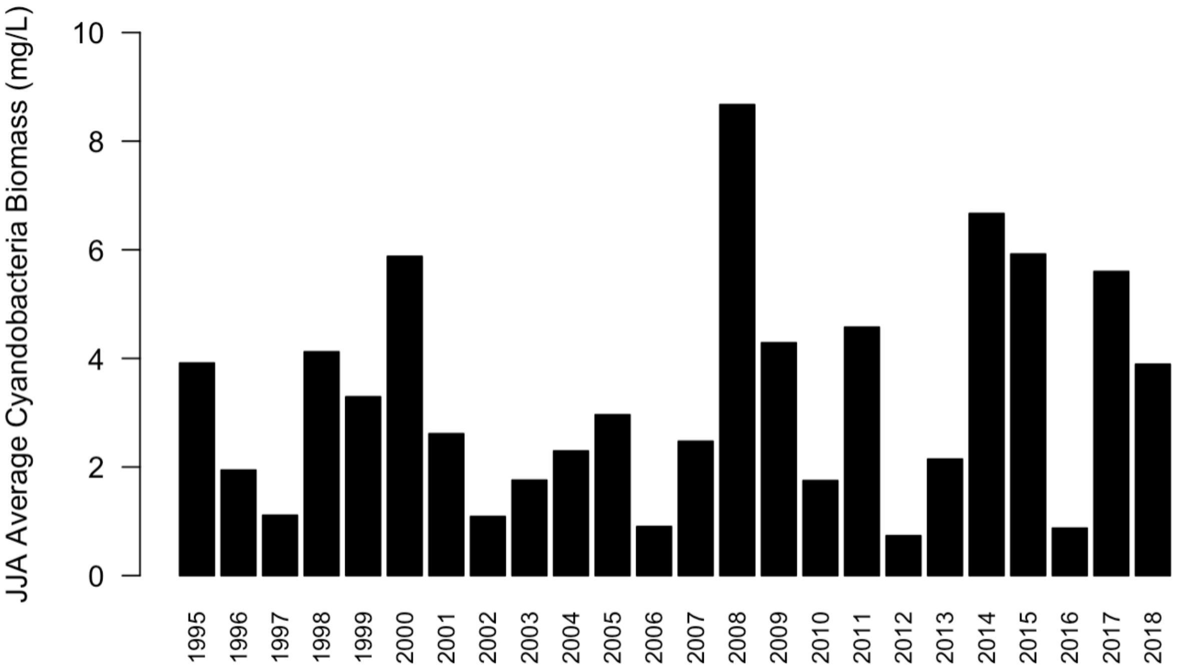

Figure 2A-1. June-August (JJA) average cyanobacteria biomass for 1995-2018 measured at the NTL-LTER buoy in Lake Mendota (see Fig. 1 for location). Aggregated from composite samples 0-8m of depth, taken at the deepest point in Lake Mendota.

2.A.2 Materials and Methods

Forecasting models for July-August cyanobacteria biomass and beach closings are built and validated from 1995 to 2018 (24 years) and 2005 to 2020 (16 years) respectively. In the following *Local scale* and *Global scale* sections a literature review is conducted to identify potential pre-season drivers of summertime cyanobacteria abundance. Predictors should be based on readily available preseason (March-May, MAM) observations to facilitate real-time predictions and must be significantly correlated (95% confidence level) with June-August (JJA) cyanobacteria biomass and beach closings. *Model construction* describes a principal component regression modelling approach, and metrics to quantify model skill are defined in *Model Performance Metrics*.

Phytoplankton samples are taken in Lake Mendota using a tube sampler in the deep hole region of the lake. Samples are collected as a composite whole water sample from 0 – 8m of depth. Phytoplankton biovolume is measured by PhycoTech, inc. (Magnuson et al., 2022). To compute biomass, biovolume was initially calculated for each species by multiplying the average cell volume for the geometric solid by the cell density in the water sample and then converting mm^3/mL of biovolume to mg/L of biomass. To describe seasonal beach closings, two separate metrics were developed: beach days closed (number of days a beach is closed during a single JJA season due to cyanobacteria, Figure 2A-2), and beach periods closed (number of periods a beach is closed during a single JJA season, defined as one or more consecutive days closed, Figure 2A-3). Together, these two metrics better define the distribution of beach closings across the season by detailing the total number of days closed and how those days are grouped throughout the season. The suite of potential predictors includes persistent large-scale climate variables and local spring drivers of cyanobacteria. Similar predictors were assessed for the cyanobacteria biomass model and beach closing model, however, both models were not required to retain the same set of predictors.

Expanding the suite of predictors beyond springtime phosphorus allows for evaluation against a previous season-ahead prediction cyanobacteria prediction model on Lake Mendota developed by Stow et al. (1997).

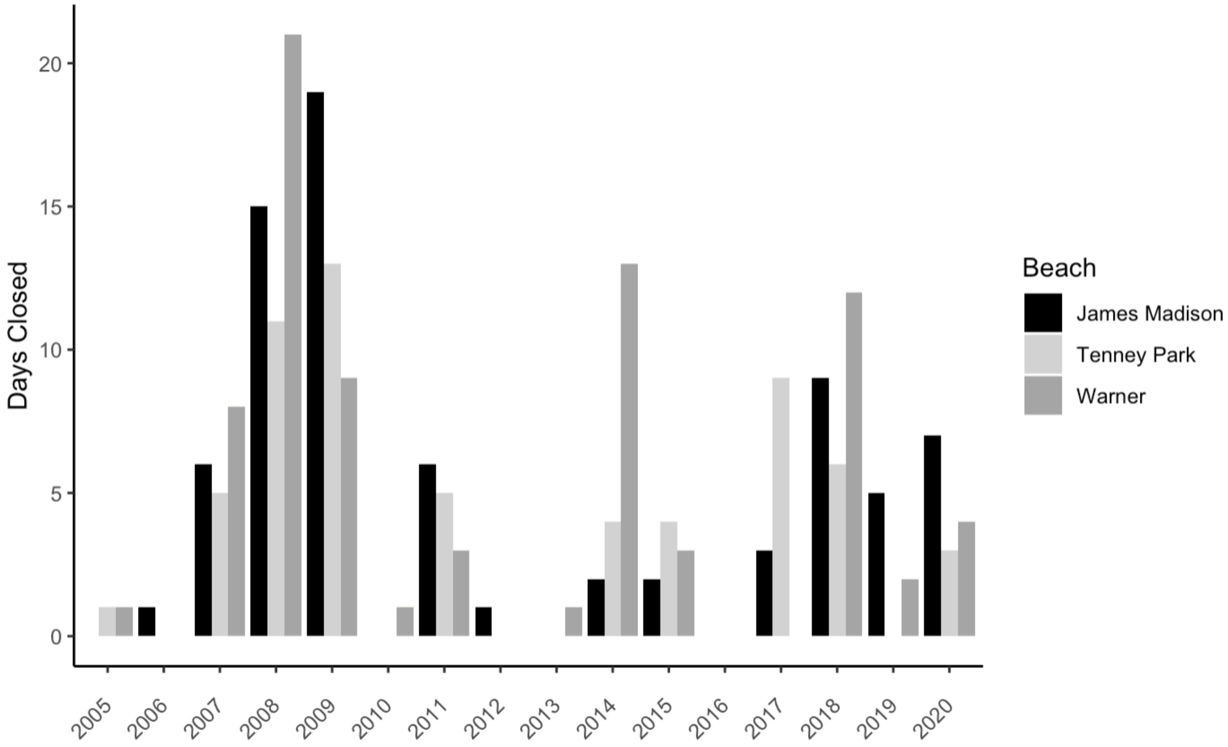

Figure 2A-2. June-August (JJA) beach days closed due to cyanobacteria abundance (data courtesy of Madison-Dane County Public Health) for 2005-2020.

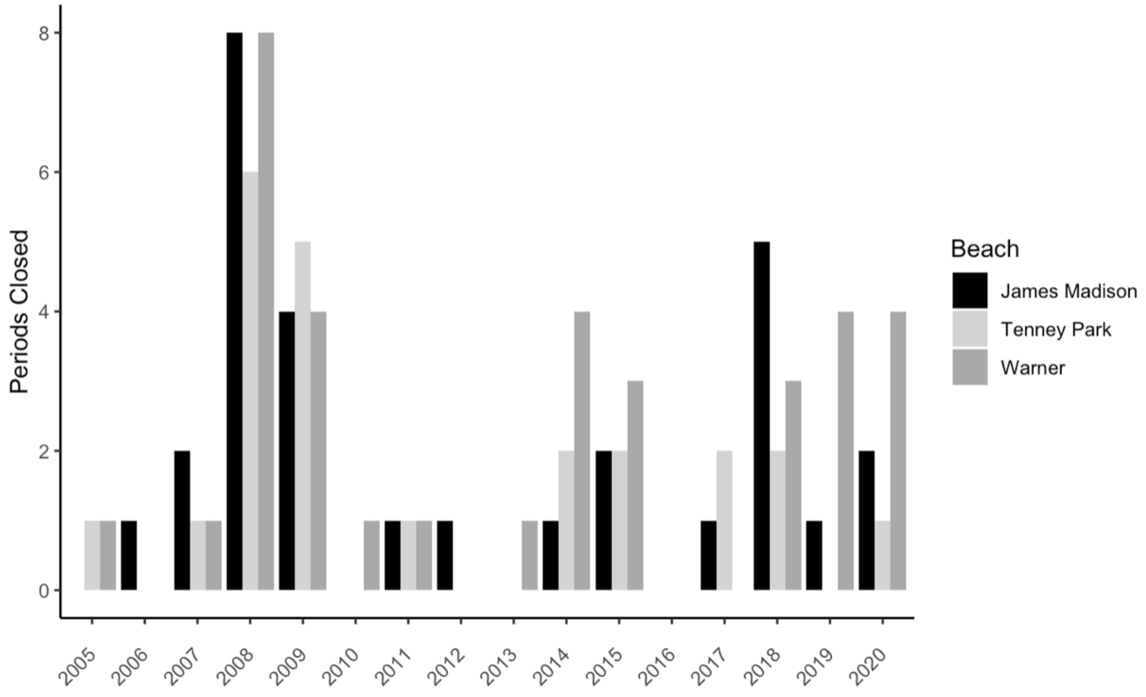

Figure 2A-3. June-August (JJA) beach periods closed due to cyanobacteria abundance (data courtesy of Madison-Dane County Public Health) for 2005-2020.

2.A.2.1 Local scale

Prospective local-scale spring drivers include residual (legacy) and external phosphorus loadings and meteorological variables, such as temperature and precipitation. Phosphorus is recognized as the driving nutrient for primary production in many lake ecosystems (Bennett et al. 1999, Paerl 2017), and the relationship between algal biomass and total phosphorus in the growing season is well established (V. H. Smith, 1982; Vollenweider, 1971). Specifically, existing prediction models for Lake Mendota have illustrated the predictive power of spring phosphorus concentrations on summer algae abundance (R C Lathrop et al., 1998; C A Stow et al., 1997). Spring phosphorus concentrations have also been used to predict algae abundance in other temperate waterbodies (Dillon & Rigler, 1974; Obenour et al., 2014; Stumpf et al., 2012; Stumpf, Johnson, et al., 2016)

Additionally, researchers have noted the influence of meteorological variables on cyanobacteria abundance, including springtime temperature and precipitation (Paerl & Huisman, 2008; Reichwaldt & Ghadouani, 2012; Craig A Stow et al., 2015).

Numerous studies have demonstrated that phosphorus and nitrogen are major limiting nutrients for algal growth in inland lake ecosystems (Carey et al., 2012; Edmondson & Lehman, 1981; Paerl, 2017), thus springtime phosphorus and nitrogen loads were evaluated as potential predictors of summertime cyanobacteria abundance. Precipitation and discharge during the spring season are thought to impact cyanobacteria abundance through the conveyance of nutrients from the watershed. Large precipitation events can flush high concentrations of nutrients into lakes, spurring algae growth (Stephen R Carpenter et al., 2018; Schueler, 1987). Higher intensity storms increase discharge, which tend to transfer higher concentrations of nutrients than lower intensity storms and their associated flows (Reichwaldt & Ghadouani, 2012).The intensity and frequency of springtime precipitation events affect the discharge loading concentration, distribution, and residence time of phosphorus within a lake, and further influence the overall availability of nutrients to cyanobacteria in the summer season (Paerl & Otten, 2016; Reichwaldt & Ghadouani, 2012; Craig A Stow et al., 2015). Therefore, total precipitation, extreme precipitation events (>40 mm/day) and discharge from March-May were considered as potential predictors of summertime cyanobacteria abundance. While total precipitation and number of extreme events are similar predictors, they represent distinct hydrologic phenomena. Total precipitation may better represent moisture conditions in the watershed compared to extreme precipitation which may lead to large runoff and nutrient loading events. Precipitation data are taken from the Midwest Regional Climate

Center and phosphorus and discharge data are taken from the United States Geological Survey (USGS gages 05427718 and 05427850) (Survey, 2021a, 2021b; Wuertz et al., 2018).

2.A.2.2 Global scale

Large-scale atmospheric-oceanic climate variables may influence local cyanobacteria abundance through atmospheric teleconnections, which influence meteorological conditions over the watershed from year-to-year. Global sea-surface temperatures (SST) and sea level pressures (SLP) are representative of these teleconnections and are well-established as drivers of precipitation and temperature on seasonal timescales by altering atmospheric flow (Markowski & North, 2003; Trenberth & Caron, 2000). Therefore, regions of SSTs and SLPs are examined as potential predictors. The El Niño Southern Oscillation (ENSO), an anomalous warming or cooling of SST in the equatorial Pacific Ocean, is perhaps the most well-known and studied oceanic-atmospheric climate phenomena with global impacts (Ropelewski & Halpert, 1986, 1987; Sarachik & Cane, 2010). In the upper Midwest ENSO is associated with warmer and drier winters during El Nino phases (Center, 2016; Impacts, 2011; Legler et al., 1999; S. R. Smith et al., 1999), contributing to lower antecedent soil moisture conditions. Although the summertime influence of ENSO in the Midwest is less pronounced, early summer months have been characterized as cooler and wetter than normal in El Nino years establishing conditions for higher runoff and nutrient transport potential. Both global and ENSO-related SST predictors were therefore considered as predictors and were identified using gridded correlation maps. SST data is retrieved from the NCEP/NCAR reanalysis (NCEP, 1994). In addition to selecting regions that meet the 95% statistical significance level requirement, distinct teleconnections between oceanic-atmospheric regions and the upper Midwest U.S. must also exist. SSTs are particularly advantageous from a prediction perspective as they fluctuate slowly over time, often allowing anomalies to persist across seasons. Similarly, sea level pressure is also evaluated globally.

2.A.2.3 Model Construction

A principal component analysis (PCA) and regression modeling approach was selected to predict cyanobacteria abundance and beach closings. PCA decomposes a space-time random field – all potential season-ahead predictors in this case – and produces a set of orthogonal time patterns that include the dominant signals, or principal components (PCs), stemming from the original set of predictors (Block et al., 2009; Von Storch & Zwiers, 2002). Additionally, PCA efficiently accounts for multi-collinearity that may be present in the predictors, a common problem in linear regression. Typically, the first few PCs explain the majority of variance in the data. Kaiser's Rule was adopted, which specifies retaining all PCs with eigenvalues greater than one (Kaiser, 1960). The retained PCs are then applied as predictors in a multiple linear regression model to predict JJA average cyanobacteria biomass and beach closings (independently). Leave-one-out cross-validation was applied for a hindcast assessment across 1995-2017 to evaluate model skill. This PCA leave-one-out cross-validation model takes the general form of Equation 1, where α_i is a fitted value, PCj_i represents the *j*-th principal component calculated with the i-th year dropped, and βj_i is the fitted coefficient for the *j*-th principal component, and $\hat{Y}_i$ represents the predicted value for the i-th year. To account for uncertainty, random deviates from the standard deviation of the prediction error are added to the model (median) prediction (Helsel & Hirsch, 1992).

$$\hat{Y}_i = \alpha_i + \beta 1_i PC1_i \; + \beta 2_i PC2_i \ldots + \beta j_i PCj_i \tag{1}$$

2.A.2.4 Model Performance Measures

To assess model performance, model results were compared with observations of cyanobacteria biomass and beach closings using three performance measures: Heidke Skill Score (HSS), Ranked Probability Skill Score (RPSS), and a Hit-Miss Matrix. Pearson and Spearman correlation coefficients, Forecast Bias, and False Alarm Ratio (FAR) were also calculated for the cyanobacteria biomass forecast. Forecast bias is the ratio of how often a specific category is forecasted to how often the specific category is observed with a value equal to one indicating an unbiased forecast and values greater than and less than one indicating over-forecasting and under-forecasting, respectively (Dee & Da Silva, 1998). FAR is a simple ratio of the number of non-occurrence forecasts of a specific category and the total number of times the specific category is forecasted. Values for this metric range from 0 to 1, where 0 indicates a perfect score (Schaefer, 1990).

Both HSS (Equation 1) and RPSS (Equation 2) report the model's ability to predict categorical outputs (e.g. high vs. low) compared to a reference forecast, typically based on observed data (climatology). For hydro-climate prediction, a three-category division is often adopted, with the reference forecast based on equal probability of categories (33% each) (Alexander et al., 2019; Block et al., 2009; Lala et al., 2020). Here, the reference forecast is split into three categories of equal probability (33% each), representing *below normal* (0 - 2.18 mg/L)*, near normal* (2.18 - 4.07 mg/L)*,* and *above normal* (4.07+ mg/L) cyanobacteria conditions, denoted as [B N A]. For beach days closed and beach periods closed, a two-category division with *normal* (x(*i*)≤ mean(closed)) and *above normal* (x(*i*)> mean(closed)) was adopted and denoted as [N A], where x represents the observed number of beach days closed in the *i*-th year. The observational probabilities of each category are not equal in this case and are unique to each beach location.

The HSS takes the general form of Equation 2, which describes forecast skill in terms of i, j= [B N A]. The joint distribution of forecasts and observations is described by $P(F_i, O_j)$ while the marginal distributions of forecasts and observations are described by $P(F_i)$ and $P(O_j)$, respectively (Wilks 2011). HSS values range from -∞ to 1, where 0 represents no improvement over the reference forecast (climatology) and 1 represents a perfect forecast.

$$\text{HSS}=\frac{\sum_i P(F_i,O_j)\ -\ \sum_i P(F_i)P(O_i)}{1-\sum_i P(F_i)P(O_i)} \tag{2}$$

The RPSS measures forecast skill by accounting for the magnitude of error in the forecast, differentiating from HSS (Wilks 2011). For example, in the case of a [B N A] category forecast, if the *above normal* category is observed, RPSS would penalize a forecast that predicts *below normal* conditions more than a forecast that predicts *near normal* conditions. First, the ranked probability score (RPS) is calculated according to Equation 3:

$$RPS=\frac{1}{n_{cat}-1}\sum_{i_{cat}}^{n_{cat}}(Pcumfct_{i_{cat}}-Pcumobs_{i_{cat}})^2 \tag{3}$$

where n_{cat} is the number of forecast categories and i_{cat} is the category number. $Pcumfct_{icat}$ and $Pcumobs_{icat}$ are the cumulative probability vectors of the forecast and observation, respectively, for the specific category of interest. RPSS then compares the RPS of the forecast, RPS_{fct}, to the RPS of climatology, $RPS_{clim,}$ using Equation 4:

$$RPSS=1-\frac{RPS_{fct}}{RPS_{clim}} \tag{4}$$

As with HSS, RPSS values range from -∞ to 1, where 0 represents no skill and 1 represents a perfect forecast.

2.A.3 Results

2.A.3.1 Cyanobacteria Biomass Model

As previously detailed, both local and global scale variables are considered as potential predictors for JJA average cyanobacteria biomass. Local scale predictor variables meeting the established criteria include March - May phosphorus loadings and discharge from the Yahara River at the mouth of Lake Mendota, total April precipitation, and precipitation events exceeding 40 mm per day from the Madison-Dane County regional airport (Table 2A-1).

Table 2A-1. Pearson and Spearman correlation coefficients between June-August (JJA) average cyanobacteria biomass and March-May (MAM) potential predictor variables (1995-1996, 1998-2017); asterisks indicate statistical significance at the 95% level (1 = 1995-2002 interpolated from upstream USGS station 05427718 2 = 1995-2008 interpolated from upstream USGS station 05427718)

Cyanobacteria Biomass Predictors	Pearson	Spearman
MAM Precipitation Events > 40mm per day (MRCC)	0.58*	0.56*
April Total Precipitation (MRCC)	0.46*	0.44*
MAM Avg. Discharge (USGS Station 05427850)^1	0.42*	0.39
MAM Avg. SST in Equatorial Pacific (NOAA)	-0.44*	-0.45*
MAM Avg. External Phosphorus Load (USGS Station 05427850)^2	0.38	0.48*

SST in the equatorial Pacific Ocean correlate strongly with Mendota's summertime (JJA) cyanobacteria biomass (Figure 2A-4), a region typically associated with ENSO. Xiao et al. (2019) found evidence for synchronization between phytoplankton dynamics and ENSO in northern Wisconsin lakes, suggesting that ENSO has some influence on local climatic conditions. Although other oceanic regions of statistically significant correlation between SST and summertime cyanobacteria abundance exist (Figure 2A-4), teleconnections between these regions and the upper

Midwest are not overly apparent, therefore the selection of SST is restricted to the equatorial Pacific Ocean.

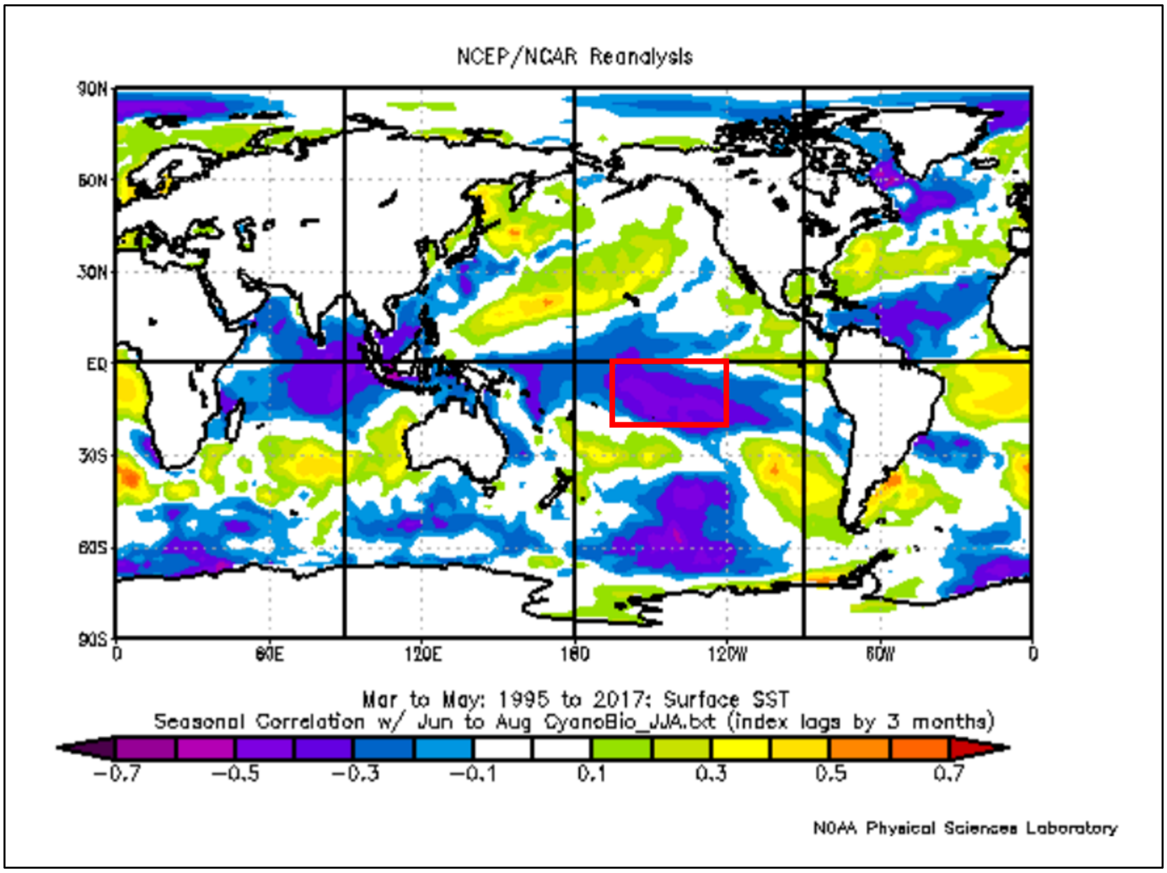
NCEP/NCAR Reanalysis
90N
60N
30N
EQ
30S
60S
90S
0
60E
120E
180
120W
60W
0
Mar to May: 1995 to 2017: Surface SST
Seasonal Correlation w/ Jun to Aug CyanoBio_JJA.txt (index lags by 3 months)
−0.7
−0.5
−0.3
−0.1
0.1
0.3
0.5
0.7
NOAA Physical Sciences Laboratory

Figure 2A-4. Correlation map of March-May (MAM) average SSTs and July-August (JJA) average cyanobacteria biomass; MAM average SSTs in the red box (190W-120W, 0-20S) are selected as a potential predictor.

Although phosphorus load meets the inclusion criteria for model development (correlation at the 95% confidence level), higher forecast skill is achieved without including phosphorus in the final suite of predictors. Thus, the final suite of season-ahead (March-May) predictors includes average discharge, the number of extreme precipitation events, total April precipitation, and average SST in the equatorial Pacific Ocean. According to Kaiser's Rule, only the first PC, explaining approximately 45% of the variance, is retained for inclusion in the prediction model. A cross-validated hindcast produces a Pearson correlation coefficient of 0.62 between median model outputs and observed cyanobacteria biomass, indicating moderate predictive skill (Figure 2A-5). This marks an improvement on previous Lake Mendota models. For example, the model developed by Stow et al. (1997) using spring center-of-lake phosphorous as a predictor of summertime cyanobacteria – with the addition of data from 1995 - 2017– has a cross validated Pearson correlation coefficient of 0.46.

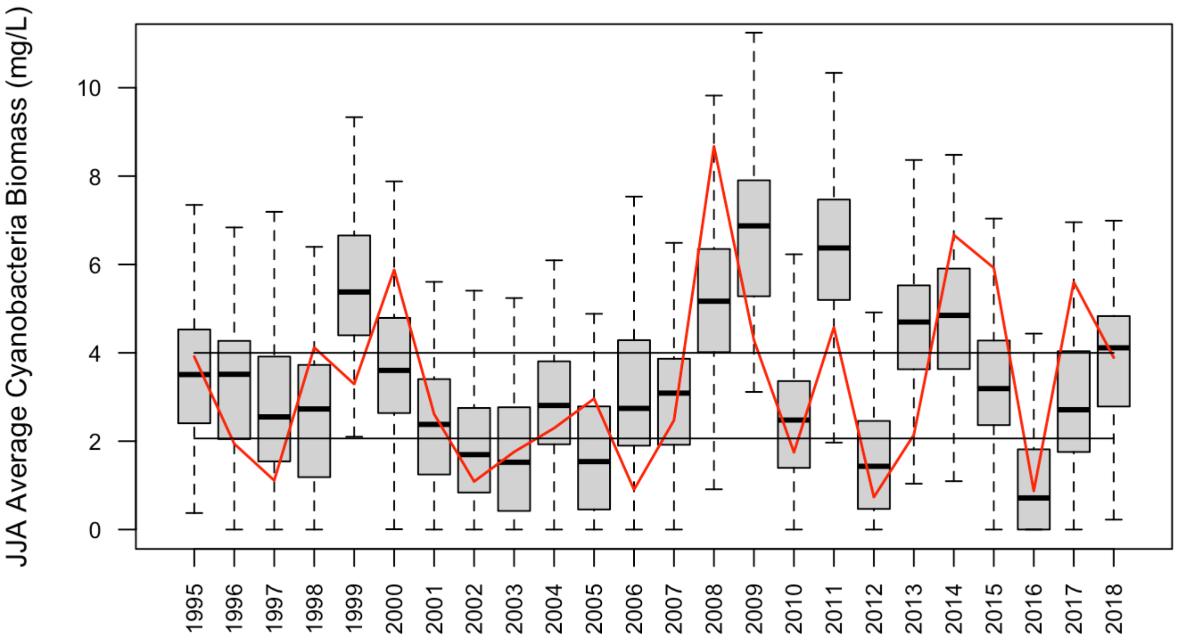

Figure 2A-5. Time-series of June-August (JJA) average cyanobacteria biomass observations (red line) and predictions (boxes); categories separated by solid black lines. Composite phytoplankton samples 0-8m of depth (Magnuson et al., 2022).

The RPSS and HSS values based on categories of equal probability are 0.60 and 0.38, respectively, indicating improvement over climatology, and model ability to generally shift toward the appropriate category. The Hit-Miss matrix (Table 2A-2) based on the [B N A] categorical divisions demonstrates high agreement, however, there is a slight propensity toward predicting *near normal* conditions when *above* and *below normal* conditions are observed. Additionally, the Hit-Miss matrix (Table 2A-2), FAR and Forecast Bias (Table 2A-3) all suggest that the model is slightly biased towards the near normal category. The model's ability to skillfully predict the *above normal* category – when cyanobacteria is most abundant and managers most concerned – is highly advantageous, however, the cyanobacteria peaks in 2008 and 2017 are clearly under predicted. Both underpredictions may be related to the distribution of precipitation throughout the spring. While both 2008 and 2017 had high overall precipitation, many days did not actually surpass the 40mm per day threshold and were thus not counted. Additionally, cyanobacteria biomass is substantially over predicted in 2009 and 2013 even though the model average prediction is still in the appropriate category. Both years saw a relatively high number of extreme precipitation events and increased streamflow. Furthermore, a limitation of the hydro-climatic forecasting approach for water quality variables is the difficulty in capturing food web dynamics, which play a significant role in structuring cyanobacteria communities in Lake Mendota (Kasprzak & Lathrop, 1997; Walsh et al., 2017). Shifts in food web dynamics may have an influence on summertime cyanobacteria abundance that is not captured in the model.

Table 2A-2 Hit-Miss Matrix for categorical June-August (JJA) average cyanobacteria biomass prediction and observations. (B = Below Normal, N = Normal, A = Above Normal).

		Forecast		
		B	**N**	**A**
Observed	**B**	5	3	0
	N	2	4	2
	A	0	3	5

Table 2A-3. Forecast bias and false alarm ratios for categorical JJA average cyanobacteria biomass predictions. (B = Below Normal, N = Normal, A = Above Normal).

Category	Forecast Bias	False Alarm Ratio (FAR)
B	0.88	0.29
N	1.25	0.6
A	0.88	0.29

2.A.3.2 Beach Closings

Categorical forecast models for beach days closed and periods closed are developed for three beaches located along the eastern side of Lake Mendota (Fig. 1). Selected season-ahead predictors mirror those included in the cyanobacteria biomass model, including average discharge, P loading, the number of extreme precipitation events, and Pacific Ocean SST, however, positive correlations between number of days or periods closed and average discharge and extreme precipitation events were the only significant correlations at any of the beaches (Table 2A-4). As with the cyanobacteria biomass model, only the first PC is retained for inclusion in each of the beach prediction models. Cross-validated hindcast model results for days and periods closed at each beach indicate moderate to strong model skill and an improvement over climatology in most metrics (Table 2A-5, Table 2A-6 for James Madison only).

Table 2A-4 Pearson correlation coefficients between June-August (JJA) beach days/periods closed and March-May (MAM) potential predictor variables; asterisks indicate statistical significance at the 95% confidence level.

Beach Closing Predictors	Characteristic Predicted	J.M.	Tenney	Warner
MAM Precipitation Events > 40mm per day (MRCC)	Days Closed	0.75*	0.70*	0.69*
April Precipitation (MRCC)		0.21	0.41	0.37
MAM Avg. Discharge (USGS Station 05427850)		0.65*	0.47	0.3
MAM Avg. SST in Equatorial Pacific (NOAA)		-0.42	-0.43	-0.46
MAM Avg. External Phosphorus Load (USGS Station 05427850)		0.39	0.32	0.39
MAM Precipitation Events > 40mm per day (MRCC)	Periods Closed	0.69*	0.87*	0.81*
April Precipitation (MRCC)		0.29	0.42	0.26
MAM Avg. Discharge (USGS Station 05427850)		0.37	0.44	0.57*
MAM Avg. SST in Equatorial Pacific (NOAA)		-0.42	-0.45	-0.36
MAM Avg. External Phosphorus Load (USGS Station 05427850)		0.16	0.39	0.36

Table 2A-5 Ranked Probability Skill Scores (RPSS), Heidke Skill Scores (HSS), and Pearson correlations for beach days and periods closed prediction models for three Lake Mendota beaches.

	Beach Days Closed			*Beach Periods Closed*		
Beach	Median RPSS	HSS	Pearson Correlation	Median RPSS	HSS	Pearson Correlation
James Madison	0.81	0.49	0.65	0.25	0.35	0.36
Tenney	0.08	0.13	0.58	0.38	0.35	0.64
Warner	-0.01	0.21	0.31	0.69	0.49	0.66

Table 2A-6 Hit Miss Matrix for categorical beach days closed predictions and observations at James Madison beach. (N = Normal, A = Above Normal).

		Forecast	
		N	**A**
Observed	**N**	6	3
	A	3	4

Beach days closed tends to be more skillful than beach periods closed, however, performance metrics are highly sensitive to the short hindcast period and strongly influenced by data in single years. This limited number of data points is especially problematic for prediction in extreme years (e.g., 2013, Figure 2A-6). In the case of James Madison beach, predictors co-vary closely with the number of beach days closed except for 2013-2014 (Figure 2A-7). The days closed forecast results mirror the cyanobacteria abundance prediction in 2013, in that both were over predicted, potentially resulting from changes in the food web not captured by the model. The model also over predicts 2014, likely due to the elevated phosphorus levels, however this may not have materialized in beach closures due to abnormally low discharge. Additionally, 2015 and 2017 above average cyanobacteria abundance did not directly translate into above average beach closures. There are several factors that may be at play in this disconnect. Wind conditions, for example, have been shown to influence horizontal movement of surface algae (Jiancai Deng et al., 2016), which may cause blooms to concentrate away from beaches allowing them to stay open. In looking at specific years, there is expected variability due to the complex dynamics of this lacustrine ecosystem.

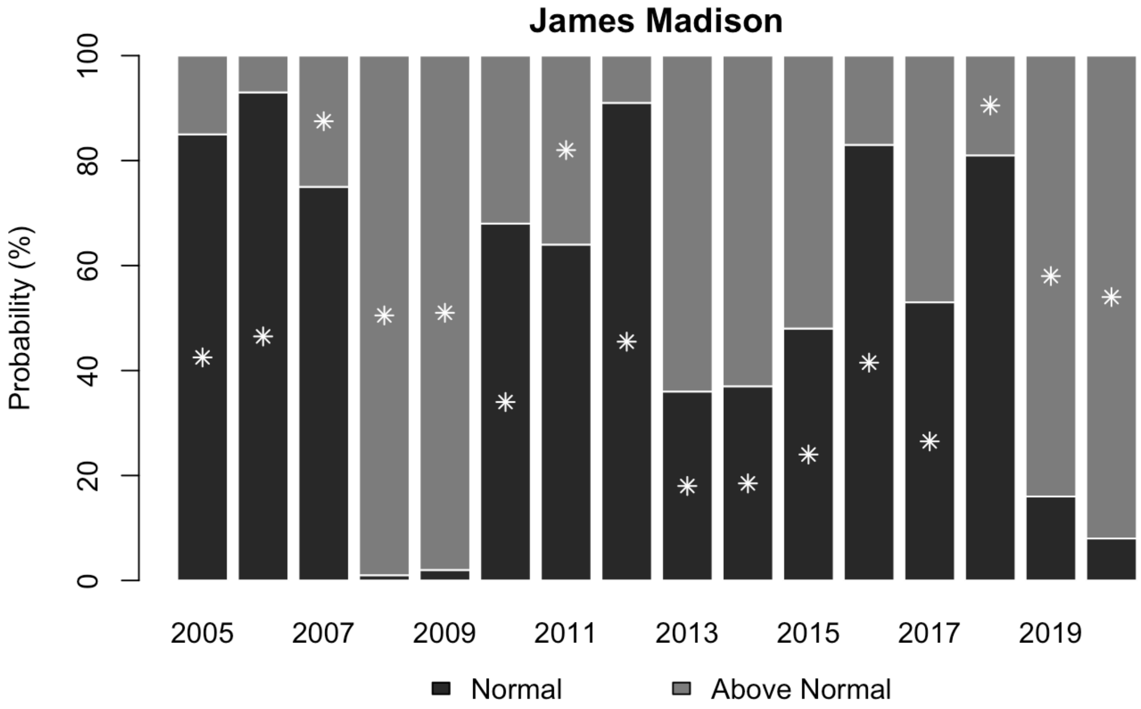

Figure 2A-6. Bar chart representing probabilistic predictions of beach days closed at James Madison beach. Normal category includes two or fewer days closed; above normal refers to more than two days closed per summer. The observed category is illustrated with a white star.

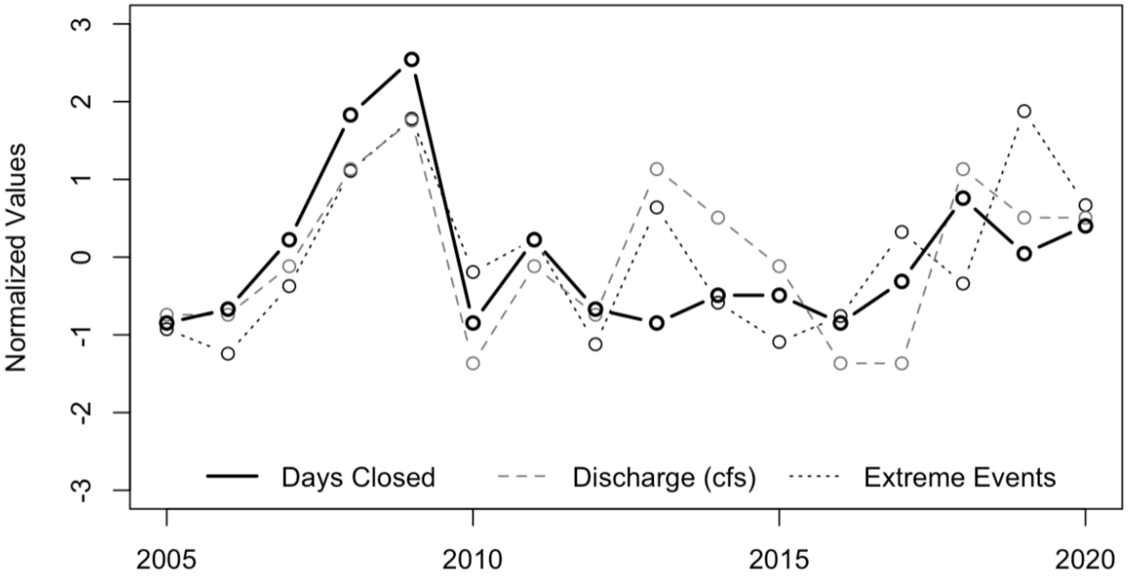

Figure 2A-7. Time-series of normalized predictors and beach days closed for James Madison beach.

2.A.4 Discussion

The development and evaluation of prospective season-ahead prediction models for cyanobacteria biomass and beach closures, based on local and global scale predictors, are presented. The model is developed as part of a sub seasonal to seasonal cyanobacteria forecasting system for Lake Mendota. Previous season-ahead prediction models have utilized phosphorus as the primary predictor variable, given its influence on cyanobacteria abundance and ability to be managed. Here, alternative predictors are also evaluated to better understand their potential contribution to prediction skill and ability to represent signals of phosphorus conveyance and distribution. In addition, models contingent solely on phosphorus data collection are subject to continuous sampling and processing lag times – often well beyond one season – which may serve as a major lake management disincentive. The modeling framework proposed here alleviates such dependence, demonstrating strong prediction skill. The proposed framework incorporates a larger suite of predictor variables than utilized in previous forecasts, however, the modeling approach remains straightforward – a clear strength for future applications. It should be noted that Lake Mendota has a wealth of high-quality, long-term data, which is uncommon among similar small inland lakes. Development of season-ahead forecasts for algae may benefit management practices in other lakes. While it is unlikely that nutrient loading and within-lake predictors will be as well characterized for other lakes, the hydroclimatic drivers evaluated here (e.g., precipitation, extreme precipitation events, air temperature, and sea surface temperatures) are widely available across the U.S. and may be used for forecasting applications in other lakes.

Although model performance exhibits predictive skill for cyanobacteria biomass, beach days closed, and beach periods closed, there are several noteworthy challenges. The statistical forecast models developed here are limited by the short time series available, with some inconsistencies in the ability to predict extremes. This may be addressed through calibrated physically based lake process models run in a predictive mode, potentially capturing complex dynamics across biological, chemical, and environmental processes, however, preliminary exploration has indicated poor to marginal skill for Lake Mendota. Another remaining challenge is when one category is predicted with high probability (confidence), yet observations fall in a different category (e.g., 2013, Figure 2-7). This is different than moderate probability of being in the unobserved category and may be a challenge to resource managers. Related, the thresholds between categories utilized here are subjective, however selection does impact model performance. Individual managers are likely to have their own preferred thresholds, warranting further evaluation into model performance for specific choices.

As discussed previously, both temperature and phosphorus load are well-established as drivers of cyanobacteria productivity, however, neither variable added predictive power at the seasonal scale. The Stow et al. (1997) prediction model uses April within-lake phosphorus concentrations to predict July-September cyanobacteria abundance with notable skill. Additionally, in the complementary sub-seasonal forecasting model for cyanobacteria abundance, June external phosphorus loads were highly correlated with July-August cyanobacteria abundance (Beal et al., 2021). It is possible that the temporal difference between the phosphorus predictor and summertime cyanobacteria biomass is responsible for this difference in skill. Internal phosphorus loading is also a significant source of phosphorus for Lake Mendota during in the summer and is not accounted for in this set of predictors (P A Soranno et al., 1997).

Spring air temperatures have been shown to influence water temperature and summertime bloom onset (M. Zhang et al., 2016), prompting the inclusion of spring air temperature as a potential predictor of summertime cyanobacteria biomass. Air temperature may have direct and indirect impacts on cyanobacteria abundance (Taranu et al., 2012). High temperatures (above 25C) during the growing season generally promote cyanobacteria growth over phytoplankton taxa such as diatoms and green algae (Paerl & Huisman, 2008). Higher air temperatures may indirectly favor cyanobacteria given that increased temperatures promote stratification strength, allowing cyanobacteria to outcompete other algal groups by using specialized gas vacuoles to adjust their position in the water column (Joehnk et al., 2008; Paerl & Huisman, 2008) . Additionally, water temperatures have been shown to control summertime cyanobacteria productivity in Lake Mendota (Konopka & Brock, 1978), however, none of the temperature-based predictors investigated correlate at a statistically significant level with biomass. There may be several explanations for this: researchers have noted that higher temperatures have a direct effect on the timing and proportional dominance of cyanobacteria, but not the amount of annual biomass (J Alex Elliott, 2012; Wagner & Adrian, 2009). While a causal relationship has been demonstrated between spring air temperatures and summertime cyanobacteria abundance in subtropical regions (Jianming Deng et al., 2014; Paerl & Huisman, 2008; M. Zhang et al., 2016), that same relationship has not been shown to exist in the northern temperate climate of the study site. This could be due to the temporal mismatch of seasonal abundance with spring temperatures. While high spring temperatures may encourage cyanobacteria dominance this does not necessarily imply long term abundance (Anneville et al., 2015; Persaud et al., 2015; M. Zhang et al., 2016). Additionally, temperature fluctuations occurring in the spring are accompanied by a variety of additional environmental changes, complicating the direct cyanobacteria response to temperature (Konopka & Brock, 1978). The simplicity of the temperature-based predictors proposed may not be capable of fully capturing the summer cyanobacteria biomass response to temperature nuances occurring throughout the season. Clearly, the predictor variables considered in this study may impact individual cyanobacteria communities differently, however, average cyanobacteria biomass across all communities is specifically addressed here as current management practices do not consider the presence of individual communities. Still, there is clear merit in the consideration of individual cyanobacteria communities that pose a greater toxicity risk, specifically those that have the potential to produce toxins, for future prediction efforts.

For this analysis, predictions are issued at the end of the spring season (beginning of June). This advance notice of summertime cyanobacteria conditions provides lake and beach managers with information necessary for making proactive management decisions ahead of the peak season for cyanobacteria productivity. These decisions may include changing the frequency of water quality testing, altering training and scheduling for lifeguards, tailoring public engagement strategies, and preparing emergency resources for recreators. Working with the sub-seasonal forecast developed by Beal et al. (2021), pre- and within- (summer) season predictions are issued for cyanobacteria abundance, allowing decision makers to adapt and optimize management strategies across the peak season for cyanobacteria productivity. The model developed here is a key component of this forecasting system, providing information on expected cyanobacteria abundances before recreational use of Lake Mendota begins to increase and toxin production becomes a potential public health threat. Linking seasonal and sub-seasonal cyanobacteria forecasts informs decisions at multiple timescales, allowing for an optimized approach to cyanobacteria management. Effectively implementing this forecasting system requires improved understanding of manager needs, key decisions dates, and available actions, all themes of ongoing research to facilitate how forecasts can better be integrated into lake and beach management.

Part B. A Sub-seasonal Cyanobacteria Biomass Forecast

Adapted from: Beal, M. R. W., O'Reilly, B., Hietpas, K. R., & Block, P. (2021). Development of a sub-seasonal cyanobacteria prediction model by leveraging local and global scale predictors. Harmful Algae, 108, 102100. DOI: 10.1016/j.hal.2021.102100

2.B.1 Introduction

Recently, the potential for developing sub-seasonal (i.e. within-season) forecasts for application to water management has garnered attention (Vitart et al., 2012; Vitart & Robertson, 2018) with the intention that such forecasts could bridge the gap between seasonal and short-term time scales (Shentsis & Ben-Zvi, 1999; Vitart, 2014). A season-ahead forecast for June-August cyanobacteria biomass and beach closings is developed in the first part of this chapter (Beal et al., 2022). In conversations with lake and public health mangers, there is an expressed desire to understand how cyanobacteria abundance may be changing throughout the summer season, and if a prediction update is possible. A sub-seasonal forecast of cyanobacteria biomass may indicate if expected cyanobacteria conditions are shifting within the season, providing managers with an opportunity to change the frequency of water quality monitoring, public engagement strategies, and prepare emergency resources for recreators and drinking water facilities before the potential for cyanobacteria productivity peaks. Part B of this chapter presents a study that investigates relevant pre- and within-season local and global scale drivers of inter-annual variability in summertime cyanobacteria biomass and showcases the development and verification of a sub-seasonal forecasting framework for cyanobacteria conditions. Finally, this study explores how a sub-seasonal forecast may be effectively paired with the full season-ahead forecast for holistic lake management.

Seasonal forecasts have been produced for summertime cyanobacteria biomass on Lake Mendota since 2015 (Soley, 2016). The modeling approach for this forecast is presented earlier in this chapter. The model is used to generate probabilistic forecasts of average cyanobacteria biomass for June-August (released on June 1). As discussed in the introduction, the highest cyanobacteria biomass concentrations, however, have historically occurred in July and August, further motivating the potential utility of a sub-seasonal forecast by updating later in the season.

2.B.2 Methods

2.B.2.1 Drivers of Variability in Cyanobacteria Biomass

In Lake Mendota, cyanobacteria biomass is generally most apparent from June-August (Figure 2), with elevated levels typically observable in July and August. The existing seasonal prediction model issues a June-August average cyanobacteria biomass forecast at the beginning of June, whereas the sub-seasonal July-August prediction model proposed here focuses on the peak months, taking advantage of June observations, and issues a forecast of average July-August cyanobacteria biomass at the beginning of July.

Numerous drivers and factors at local to global scales influence inter-annual cyanobacteria productivity. From a forecasting perspective, ideal predictors include pre-season (e.g. April-June), observable hydroclimatic and landscape variables that effect the state of the lake system into July-August. Potential predictors are identified based on previous literature regarding cyanobacteria dynamics and/or correlation analysis. Predictors that correlate with July-August average cyanobacteria biomass at the 95% confidence level ($P<0.05$) are considered statistically significant and added to the suite of potential predictors.

As discussed previously, phosphorus is a well-established driver of cyanobacteria biomass. Strong correlations between phosphorus in contributing waters and cyanobacteria biomass has been demonstrated repeatedly (Downing et al., 2001; Håkanson et al., 2007; R C Lathrop et al., 1998; V. H. Smith, 1985; C A Stow et al., 1997). Phosphorus loading (pounds per day) data for the Lake Mendota case study are extracted from USGS station 05427718 located on the Yahara River at Windsor, WI for the month of June. Relatedly, local scale hydroclimatic variables influencing transport of phosphorus across the landscape prior to July are also important drivers of external phosphorous loading. These include extreme precipitation events, discharge, soil moisture, and suspended sediments (Carpenter et al., 2018; Michalak, 2016; Motew et al., 2017). Precipitation events can wash high concentrations of nutrients off the landscape and into surface waters, contributing to eutrophication (Sinha et al., 2017). Agricultural watersheds similar to the Lake Mendota watershed are particularly vulnerable to phosphorus loading driven by precipitation events (Stephen R Carpenter et al., 2018; Garnache et al., 2016). Most of the phosphorus loading in the Mendota watershed occurs in a relatively small number of large loading events (Stephen R Carpenter et al., 2015). Similarly, soil moisture conditions regulate infiltration versus direct runoff into rivers or lakes. Intense loading events occurring in June, represented by phosphorus load, discharge, and suspended sediments, have the potential to alter phosphorus availability for cyanobacteria later in the summer (Richard C Lathrop & Carpenter, 2014). Precipitation data was obtained from the Midwest Regional Climate Center for March-May (Wuertz et al., 2018). Soil moisture data in the Mendota watershed comes from the North American Land Data Assimilation System (NLDAS) for June (Mocko, 2013). Both discharge and suspended sediment loads are from USGS station 05427718 for the month of June (Survey, 2021a).

Variables related to in-situ productivity, including nitrate + nitrite and total unfiltered phosphorus are also considered. While phosphorus has historically been considered the primary limiting nutrient for phytoplankton in freshwater, there is evidence that inorganic nitrogen can control growth and toxicity of cyanobacteria as well (Gobler et al., 2016). These data are available in the LTER database for June (Magnuson et al., 2023a). To further assess the state of lake productivity in June, a Floating Algae Index (FAI) was generated using remotely sensed images in June from Landsat 5 (1995-1999) and MODIS (2000-2017) satellites (ORNL DAAC 2020; USGS 2020), using methods outlined by Hu (2009). The FAI has been used for mapping floating algae, including cyanobacteria, in lacustrine and coastal environments (Hu, 2009; Oyama et al., 2015). The effects of nutrient loading on algal biomass are well established, therefore, an estimation of algal biomass in June may indicate the general state of productivity in the lake and serve as a predictor of cyanobacteria productivity later in the summer (Vollenweider, 1971).

As discussed previously, elevated air temperature is thought to favor dominance of cyanobacteria through direct effects on photosynthetic capacity and indirect effects on competition. June air temperature and the number of events exceeding the 99th percentile of air temperatures for the climate reference period (1981-2010) are included in the suite of potential predictors (Anneville et al., 2015; Gallina et al., 2011). Daily temperature data from NOAA’s Global Historical Climatology Network was accessed through the Midwest Regional Climate Center (Menne et al., 2012). Mean pre-season water temperature, accessed through the NTL-LTER data repository, is also evaluated as a potential predictor (Magnuson et al., 2023b; Robertson, 2016).

Variable grazing rates by *Daphnia spp*. may also influence cyanobacteria biomass. In Lake Mendota, Richard C Lathrop et al. (1999) found that summer water clarity is significantly greater in years dominated by *D. pulicaria* compared to *D. mendotae*. This difference in clarity has been attributed to the ability of the larger-bodied *D. pulicaria* to significantly reduce summertime algal biomass, including cyanobacteria (Epp, 1996; Kasprzak & Lathrop, 1997; Sarnelle, 2007). Furthermore, there is evidence to suggest that summertime *Daphnia* biomasses are greater in Lake Mendota when *D. pulicaria* dominate in the spring months (Lathrop et al., 1999). Thus, April-May *D. pulicaria* biomass, measured at the LTER buoy, is included as a potential predictor of July-August cyanobacteria biomass (Magnuson et al., 2019).

Sea surface temperature (SST) and sea level pressure (SLP) anomalies have been well-documented to influence precipitation and temperature on monthly to seasonal timescales by altering atmospheric flow conditions (Barnston, 1994; Farquhar, 2010; Giannini et al., 2000; Markowski & North, 2003) (Barnston, 1994; Giannini and Kushnir, 2000; Markowski and North, 2003), and are thus also considered as potential predictors. Locations of SST and SLP influencing climate conditions in the Lake Mendota watershed are further developed below. May-June SST and SLP anomalies are from NOAA's ERSST v3b and HadSLP2r datasets, respectively, and were accessed through the IRI Data Library (Allan & Ansell, 2006; T. M. Smith et al., 2008).

2.B.2.2 Prediction Modeling Approach

Like the season-ahead model, a principal component analysis and regression approach are selected to build the July-August cyanobacteria prediction model. Again, principal components that explained more than 10% of the variance in the data are retained. Principal component regression models are fit based on the retained PCs across 1995-2017. These models also take the form of equation 1A, described in chapter 2, part A. Given the large number of candidate predictors for the sub-seasonal model, the generalized cross-validation (GCV) score is used to select the best subset from the suite of predictors and is given as,

$$(5) \qquad \mathrm{GCV} = \frac{\sum_{t=1}^{N} \frac{e_t^2}{N}}{\left(1-\frac{m}{N}\right)^2}$$

where N is the number of time steps (1995-2017), m is the number of PCs (predictors) retained in each candidate model, and e_t is the residual (difference between observed and model estimated values) at each time step, t (July-August each year). The GCV penalizes overfitting and is a good estimate of predictive risk (Craven and Wahba 1978). Using this method, the best set of predictors can be identified by evaluating several predictor combinations (candidate models) and selecting the combination that results in the minimum GCV score (Regonda et al., 2006). Statistical models were developed using R version 1.3.1056.

In an effort to appropriately represent teleconnection patterns between global climate phenomenon and local-scale processes that drive cyanobacteria biomass, a Nino Index Phase Analysis (Giuliani et al., 2019a; Zimmerman et al., 2016) is adopted. This method draws on the state of the atmospheric-oceanic system in months prior to the season of interest to divide a timeseries into different “mean states”. This allows for possible asymmetric relationships between “mean states” to be captured and modeled (Lee et al., 2018). Given that ENSO expresses moderate influence over climate conditions in the upper Midwestern U.S., the Multivariate ENSO Index (MEI) – consisting of SLP and SST information in the Pacific Ocean – is used to classify historical years into phases of ENSO; here two phases are adopted: positive and negative, based on MEI values averaged over May – June. Global and local-scale predictors may subsequently be evaluated for each "mean state" of the atmospheric-oceanic system represented by each phase. For the historical cyanobacteria biomass record on Lake Mendota, seven years fall into the positive phase and 16 into the negative phase. For the years falling within each phase, regions of SST and SLP anomalies that correlated significantly with July-August cyanobacteria are selected following Zimmerman et al. (2016). Principal components of these SST and SLP regions are included as potential sub-seasonal predictors. Each of the specified predictors are evaluated independently for the positive and negative phase, as correlation with cyanobacteria biomass in one phase does not necessitate inclusion as a predictor in both phase models. Thus the important processes contributing to cyanobacteria growth in each phase of ENSO are identified, potentially leading to enhanced biomass forecasts. The Nino Index Phase Analysis was performed in Python 2.7.16 and Spyder 3.3.6 using code developed by Giuliani et al. (2019b)

To evaluate historical performance, a hindcast is undertaken, such that a year of information is dropped (drop one cross-validation), the PCs are constructed, coefficients α and β are fit based on the remaining years of data, and Y_t for the dropped year is calculated. This is repeated to create a deterministic forecast of biomass for all years. The optimal number of PCs for each model, based on the GCV, was held constant for the cross-validation in all years.

Ensemble predictions for each year in the hindcast are based on errors, defined as the difference between predicted and observed cyanobacteria biomass in the leave-one-out cross-validated approach. Errors are fit to a normal distribution, with mean zero, using a maximum likelihood estimation. For each hindcast year, 100 random draws from the distribution are added to the deterministic biomass forecast to form the ensemble (Alexander et al., 2019; Delorit et al., 2017; Helsel & Hirsch, 1992; M. Zhang et al., 2016).

2.B.2.3 Model Performance Metrics

To assess model performance, observations are compared with model forecasts using five skill scores: correlation coefficients, Root Mean Square Error (RMSE), Heidke skill score (HSS), ranked probability skill score (RPSS), and a hit-miss matrix (Heidke, 1926; Epstein, 1969). As described earlier, HSS and RPSS are categorical performance metrics and can be interpreted as a percentage improvement over a reference forecast. Here, the reference forecast is split into three categories of equal probability (33% each), representing *below normal* (0 - 2.91 mg/L)*, near normal* (2.91 - 4.56 mg/L)*,* and *above normal* (4.56+ mg/L) cyanobacteria conditions. For the forecast model developed here, if there is no predictive information, the model defaults to equal odds categorical prediction, as in the reference forecast. However, for most years, the distribution of expected conditions shifts and results in unequal likelihoods of each category. Thus, the forecast developed here outperforms the reference forecast when it assigns a greater probability (more than 33%) to the category that is ultimately observed.

2.B.3 Results

2.B.3.1 Phase Model Performance

A unique set of cyanobacteria predictors are retained for the MEI positive and negative phase models (Table 2B-1), validating the utility of separate models to describe this asymmetric relationship. Regions of May-June SST anomalies are identified following Zimmerman et al. (2016) for both the positive and negative phases (Figure 2B-1). In the negative phase (La Niña-like) model, significantly correlating regions of May-June SST anomalies are located in the equatorial Pacific Ocean. In the positive phase (El Niño-like) model, significantly correlating regions of May-June SST anomalies are located in the mid and northern Atlantic Ocean.

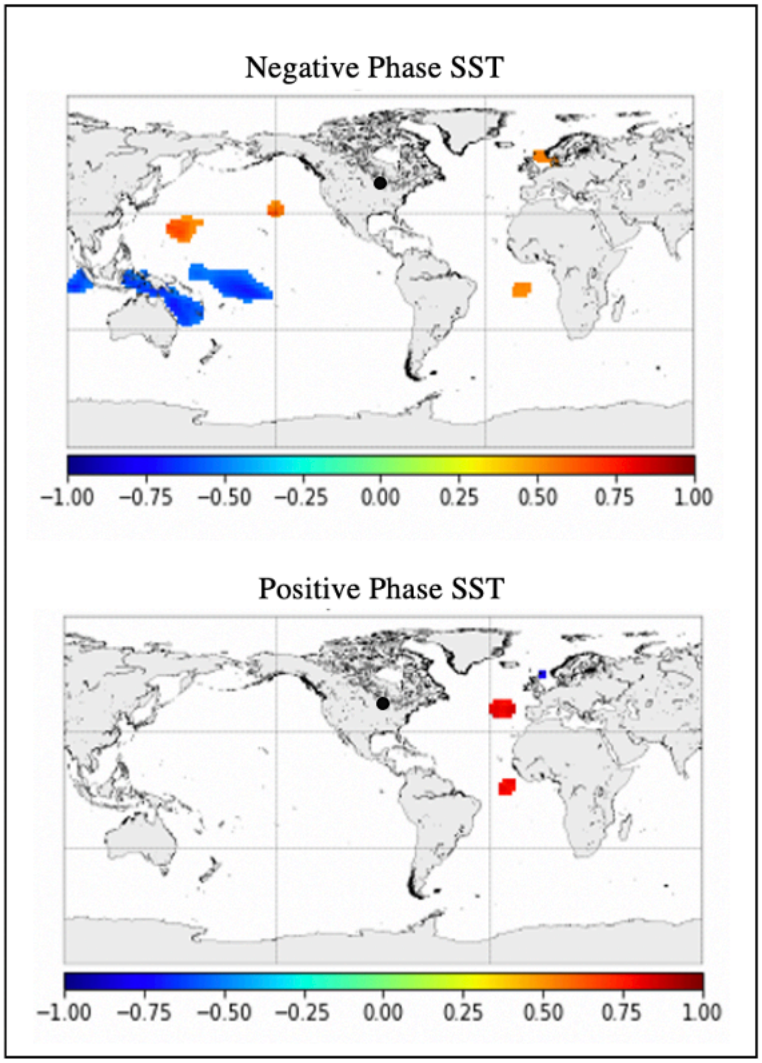

Figure 2B-1. Regions of statistically significant (95th-percentile) Pearson correlation coefficients between July-August cyanobacteria biomass and May-June SST anomalies for negative and positive ENSO phases. The black dot represents the study site. Colors represent the degree of correlation.

The final set of predictors for the negative (La Niña-like) phase includes June discharge, June phosphorus load, June total unfiltered phosphorus measured at the LTER buoy, the floating algae index for June, and May-June average SST anomalies in parts of the Pacific Ocean. These first four variables represent local-scale processes, and SSTs represent global scale processes, explaining cyanobacteria variability. The first three principal components are retained for the negative phase model and explain approximately 65%, 13%, and 10% of the variance, respectively.

The final set of predictors for the positive (El Niño-like) phase only includes May-June SST anomalies in the Atlantic Ocean and the floating algae index. The first principal component of the positive phase model explains approximately 92% of the variance in the data and is the only PC retained for the model. Variables commonly associated with cyanobacteria productivity (e.g. phosphorus, discharge, extreme precipitation events) are not statistically significant during the positive phase (Table 2B-1).

Table 2B-1. Pearson and Spearman correlation coefficients between July-August average cyanobacteria biomass with ENSO phase indicated. Bold values indicate the set of significantly correlated predictors selected with the GCV for each phase model. * indicate significantly correlated variables.

Predictor	Months	ENSO Phase	Pearson	Spearman	Source
Discharge (USGS Station 05427718)	June	Positive	-0.09	-0.04	USGS
Phosphorus Load (USGS Station 05427718)	June		0.11	-0.13	USGS
Suspended Sediment Load (USGS Station 05427718)	June		0.22	-0.11	USGS
Soil Moisture (Grid: 43.313 -89.313)	June		0.15	0.25	NLDAS
Nitrate + Nitrite (Buoy)	June		-0.41	-0.21	LTER
Total Unfiltered Phosphorus (Buoy)	June		0.21	0.04	LTER
Sea Surface Temperature (PC1)	May-June		**-0.87***	**-0.96***	IRI Data Library
Sea Level Pressure (PC1)	May-June		-0.88*	-0.96*	IRI Data Library
Extreme Events (>25mm)	March-May		0.25	0.25	MRCC
Air Temperature	June		-0.11	-0.54	MRCC
Extreme Air Temperature Events	March-June		-0.002	-0.23	MRCC
Floating Algae Index	June		**0.93***	**0.89***	MODIS/Landsat
D. Pulicaria Biomass	June		-0.89*	-0.89*	LTER
Water Temperature	April-June		0.64	0.57	LTER
Pre-season Cyanobacteria Biomass	June		0.52	0.02	LTER
Discharge (USGS Station 05427718)	June	Negative	**0.83***	**0.63***	USGS
Phosphorus Load (USGS Station 05427718)	June		**0.82***	**0.57***	USGS
Suspended Sediment Load (USGS Station 05427718)	June		0.78*	0.55*	USGS
Soil Moisture (Grid: 43.313 -89.313)	June		0.43	0.43	NLDAS
Nitrate + Nitrite (Buoy)	June		0.63*	0.66*	LTER
Total Unfiltered Phosphorus (Buoy)	June		**0.62***	**0.61***	LTER
Sea Surface Temperature (PC1)	May-June		**-0.74***	**-0.57***	IRI Data Library
Sea Level Pressure (PC1)	May-June		-0.63	-0.41	IRI Data Library
Extreme Events (>25mm)	March-May		0.72*	0.75*	MRCC
Air Temperature	June		-0.05	-0.003	MRCC
Extreme Air Temperature Events	March-June		0.09	-0.02	MRCC
Floating Algae Index	June		**0.71***	**0.70***	MODIS/Landsat
D. Pulicaria Biomass	June		-0.15	-0.03	LTER
Water Temperature	April-June		0.58	0.39	LTER
Pre-season Cyanobacteria Biomass	June		0.18	0.60	LTER

2.B.3.2 Combined Model Performance

A hindcast assessment combining the positive and negative phase models results in Pearson and Spearman correlation coefficients of 0.90 and 0.83 respectively, an RMSE of 1.22, an HSS of 0.41, and a median RPSS of 0.72, indicating a clear improvement over climatology (Figure 4). The model illustrates particular skill in predicting *below normal* and *above normal* conditions but performs poorly in the *near normal* category (Table 2B-2). The model's ability to correctly predict *above normal* July-August cyanobacteria biomass (6 out of the 8 years) is particularly advantageous from a management perspective. Additionally, the two-phase model demonstrates substantial improvement over a traditional model that does not discriminate between ENSO phases (Figure 2B-3; the "one-phase" model results in Pearson and Spearman correlation coefficients of 0.81 and 0.70, respectively, an RMSE of 1.53, an HSS of 0.35, and a median RPSS of 0.56.)

Table 2B-2 Cyanobacteria biomass forecast results: observed cyanobacteria biomass category vs. the forecasted category in a given year. Values represent the number of historical years that fall into each category based on a hindcast.

	Observed			
		Below Normal	Near Normal	Above Normal
Forecast	Below Normal	7	5	2
	Near Normal	1	1	0
	Above Normal	0	1	6

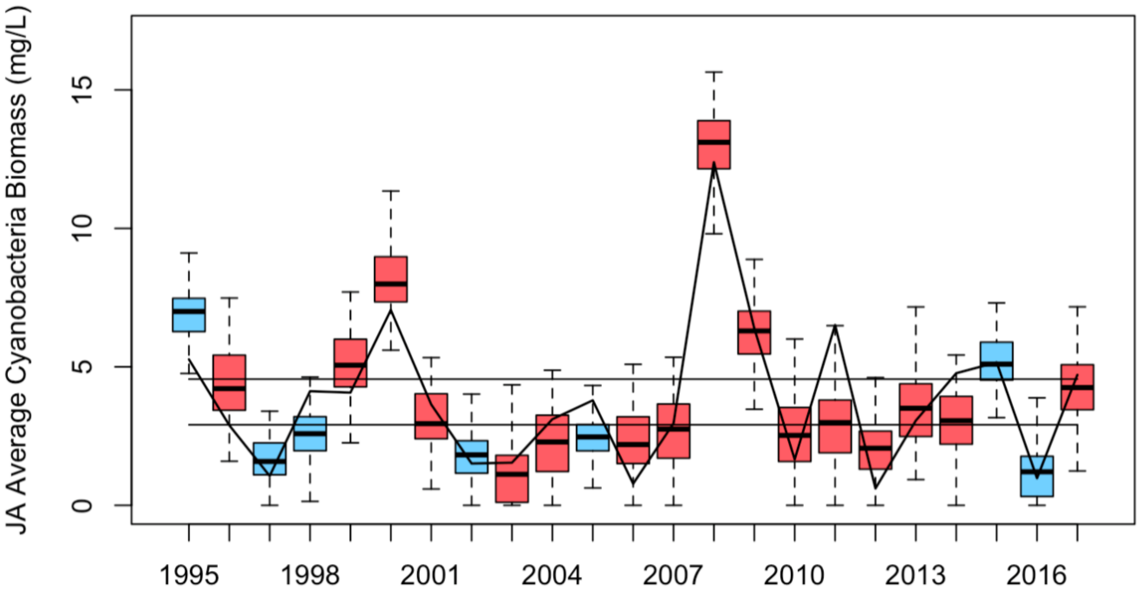

Figure 2B-2. July-August average cyanobacteria biomass predictions for positive (blue) and negative (red) phases of ENSO (box plots) and observed data (solid black line.) Thresholds between *below*, *near*, and *above* normal categories are denoted by horizontal black lines.

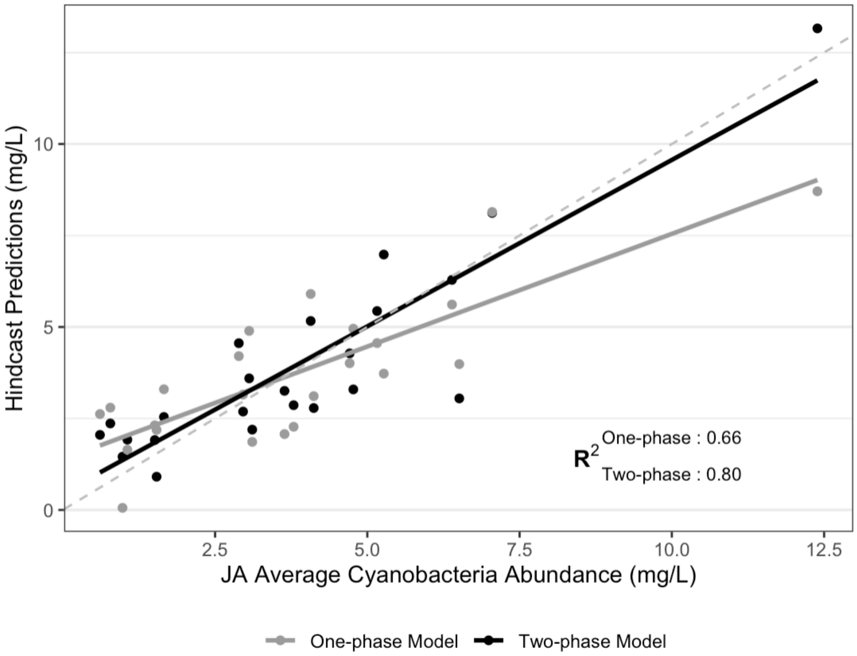

Figure 2B-3. July-August average cyanobacteria biomass observations compared to hindcast predictions using "one-phase" and "two-phase" (NIPA) models. A 1:1 line (perfect forecast) is represented by the dashed gray line.

2.B.3.3 Seasonal and Sub-seasonal Model Comparison

A comparison between the full (June-Aug) and sub-seasonal (July-Aug) model outputs is warranted to understand agreement between models and potential gains from issuing an updated forecast. A probability density plot of the sub-seasonal hindcast appears to more accurately reflect observed conditions than the seasonal hindcast and illustrates the sub-seasonal model's increased accuracy in the tails, with less emphasis on the *near normal* category (Figure 2B-4). To assess the degree of difference between the predicted probability distributions and the observed distribution, two, two-sample Komolgorov-Smirnov test are performed. The Komolgorov-Smirnov test statistic (D) quantifies the distance between two empirical distribution functions. A smaller test statistic is found between the sub-seasonal and observed distributions (D=0.22) compared to the seasonal and observed distributions (D=0.30). Neither the predicted sub-seasonal or seasonal distribution was significantly different from the observed July-August biomass distribution at the 95% confidence level (P=0.66 and P=0.23, respectively). From a categorical perspective, normalized seasonal and sub-seasonal hindcasts correctly predict 56.5% and 60.8% of observed July-August biomass respectively (Table 2B-3). Both hindcasts perform well in the *below normal* category and poorly in predicting *near normal* conditions. Most notably, the sub-seasonal prediction model for *above normal* cyanobacteria biomass is an improvement over the seasonal forecast model. Specifically, the sub-seasonal forecast correctly predicts *above normal* conditions for three years in which the seasonal forecast does not (Table 2B-4). The sub-seasonal forecast incorrectly updated an *above normal* seasonal prediction in only one year (2011). Increased accuracy in prediction of *above normal* cyanobacteria conditions by the sub-seasonal forecast is encouraging, as these conditions present the greatest threat to public health.

Table 2B-3 Observations and number of correct normalized categorical predictions of July-August cyanobacteria biomass from 1995-2017.

Category	Years observed	Correct Seasonal Forecasts (%)	Correct Sub-seasonal Forecasts (%)
Above Normal	8	3 (37.5)	5 (62.5)
Near Normal	7	2 (28.6)	2 (28.6)
Below Normal	8	8 (100)	7 (87.5)
All	23	13 (56.5)	14 (60.8)

Table 2B-4 Years in which the sub-seasonal forecast corrected an incorrect seasonal forecast (Corrections) and years in which the sub-seasonal forecast miscorrected an accurate seasonal forecast (Miscorrections). "Corrections" always imply an incorrect seasonal forecast. "Miscorrections" imply a correct seasonal forecast. Cases in which both forecasts are correct or incorrect are not represented.

Category	Corrections	Miscorrections
Above Normal	3	1
Near Normal	2	2
Below Normal	0	1

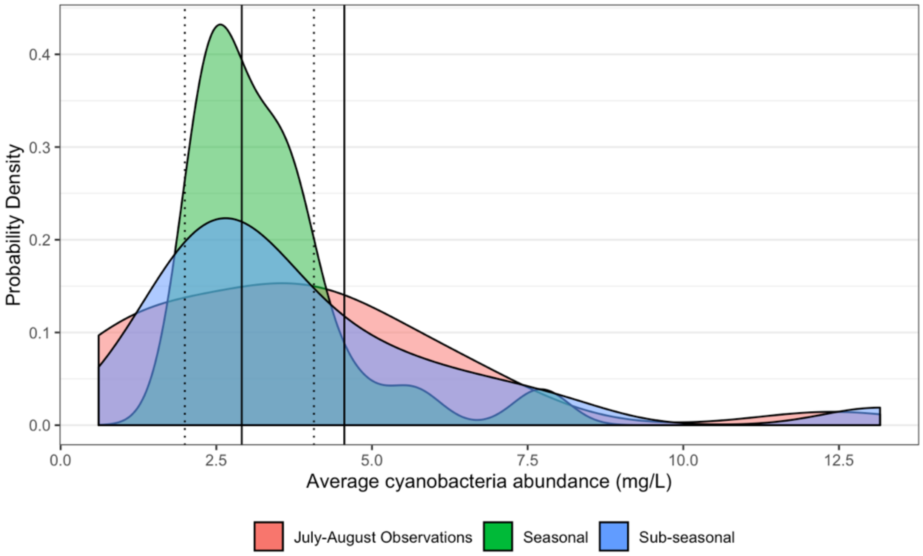

Figure 2B-4. PDFs of average cyanobacteria biomass observations (July-August), seasonal prediction (June-August), and sub-seasonal prediction (July-August across 1995-2017), represented by red, green, and blue areas, respectively. Additional colors represent areas in which the PDFs overlap. Dashed vertical lines indicate thresholds between *below, near,* and *above normal* seasonal (June-August) categories. Solid lines indicate sub-seasonal (July-August) category thresholds.

2.B.4 Discussion

In addition to practical applications, prediction plays an important role in demonstrating ecological understanding (Houlahan et al., 2017). The development and assessment of the positive and negative phase forecasting models provides some insight into the relative importance of local and global scale variables on a seasonal timeframe.

At the local scale, the predictive power of several variables are noteworthy. The floating algae index, a remotely sensed indicator of pre-season lake productivity, is the only local-scale predictor significant in both phase models. Pre-season cyanobacteria biomass, however, was not a significant predictor of July-August biomass in either phase. This suggests that general algae productivity in the early summer may be more indicative of favorable conditions for July-August cyanobacteria than early summer cyanobacteria biomass itself. Additionally, despite the established importance of temperature in cyanobacteria productivity, neither of the air temperature-based predictors are significantly correlated with July-August cyanobacteria biomass in either phase. Pre-season water temperatures resulted in higher correlation coefficients than air temperature predictors, but relationships were not strong enough to be included in either phase. Konopka and Brock (1978) purport that the relationship between lake temperature and cyanobacteria growth in Mendota is complicated by other concurrent environmental changes. Ultimately, the temperature-based predictors included here may be too simplistic to fully capture the relationships between air and water temperature and cyanobacteria growth.

At a global scale, regions of relevant sea surface temperatures identified by the NIPA process suggest differences in the influence of large-scale climate phenomena on local hydroclimatic processes in the Midwest during the positive and negative phases of ENSO. In the negative phase (La Niña-like) model, significantly correlating regions of May-June SST anomalies are located in the equatorial Pacific Ocean, a region commonly associated with ENSO (Figure 3). A relationship has been previously established between springtime La Nina conditions and a strong Great Plains low level jet (GPLLJ), which acts as a conduit for moisture from the tropical Atlantic to the continental U.S. (Munoz and Enfield, 2011; Krishnamurthy et al., 2015). Increased springtime moisture during the MEI negative phase may explain why variables associated with precipitation (e.g. phosphorus load, discharge, extreme events, suspended sediments) are significantly correlated with July-August cyanobacteria biomass in the negative phase, but not in the positive phase (Table 2B-1).

In the positive phase (El Niño-like) model, significantly correlating regions of May-June SST anomalies are located in the mid and northern Atlantic Ocean (Figure 2B-1). The GPLLJ draws moisture from the tropical Atlantic via the Caribbean low-level jet, however, Krishnamurthy et al. (2015) suggest that El Nino conditions are not typically associated with a strong GPLLJ in boreal spring (April-June). This may explain why regions of significantly correlating SST (and SLP) anomalies are focused in the Atlantic, and why they are absent from the equatorial Pacific Ocean. On average, total June precipitation is lower during the positive phase compared to the negative phase (Figure 2B-6). Furthermore, drought periods in the Yahara watershed have been shown to decrease July-August discharge, phosphorus loads, and total phosphorus within the lake (Lathrop and Carpenter, 2014). A weaker GPLLJ in the positive phase may explain lower June precipitation and the lack of significant correlations between precipitation-driven variables and cyanobacteria biomass in these years.

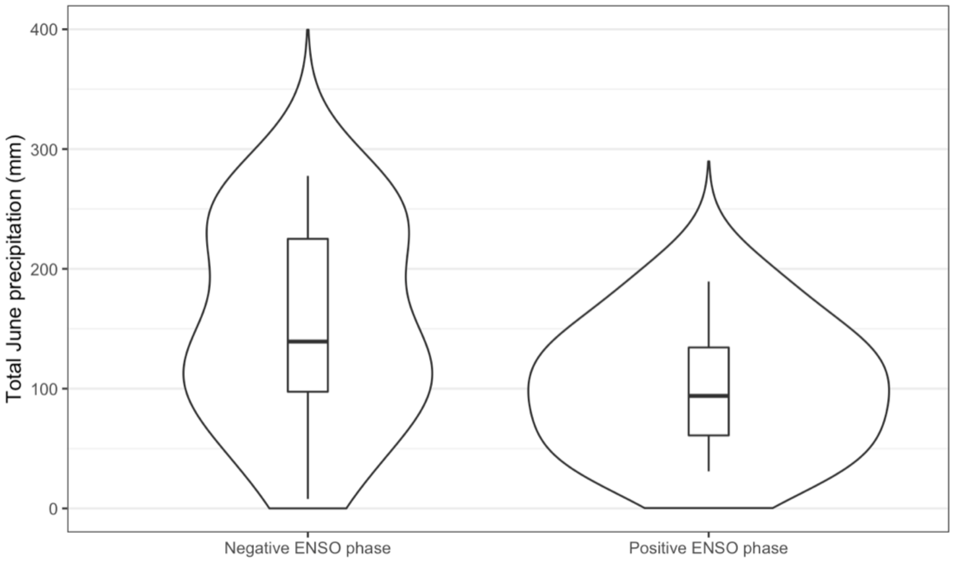

Figure 2B-5. Violin plot of total June precipitation for positive and negative ENSO phases (1995-2017.) Mean precipitation is 153.8 and 100.7 millimeters for each phase, respectively. Precipitation data recorded at the Dane County Regional Airport was obtained from the Midwest Regional Climate Center for March-May (Wuertz et al., 2020).

ENSO signals may also explain asymmetries in predictor relationships at the lake scale. *D. pulicaria* biomass is not selected by the model for the final suite of predictors but is significantly correlated with cyanobacteria biomass in the positive phase. Plankton community dynamics in Lake Mendota are complex, however, there is some evidence that the effects of *D. pulicaria* grazing on phytoplankton are more pronounced when phosphorus concentrations are low (Vanni et al., 1990). It is possible then, that the influence of *D. pulicaria* grazing on July-August cyanobacteria biomass is more pronounced in the positive phase due to differences in phosphorus conveyance between the phases of ENSO.

In comparison to a model conditioned on all available years of cyanobacteria biomass data, the split phase model approach significantly improves predictions of July-August cyanobacteria biomass, most notably for *above normal* cyanobacteria conditions (Figure 5). Additionally, the subseasonal forecast developed here shows an improvement in forecast skill for July-August cyanobacteria biomass over the full season (June-August) forecast developed by Soley et al. (2016) (Table 2B-4, Figure 2B-4). This suggests that a sub-seasonal cyanobacteria forecast (released on July 1st) can provide lake and beach managers with a meaningful update to the full season forecast (released on June 1st), with greater accuracy regarding the peak months for cyanobacteria productivity in Lake Mendota.

Despite the sub-seasonal model's ability to improve forecasts of *above normal* cyanobacteria conditions overall, the model incorrectly updated a seasonal forecast of *above normal* cyanobacteria biomass to *near normal* in 2011 (Table 2B-4). Increased cyanobacteria abundance in 2011 may have resulted from high concentrations of dissolved inorganic nitrogen throughout the summer. June-August average concentrations of total nitrogen in 2011 were the highest in the available timeseries at 1391 μg/L, notably higher than the long-term summertime average of 1156 μg/L. Nutrients concentrations are averaged across the water column, which is sampled to 25m in meter increments (Magnuson et al., 2023a). Beversdorf et al. (2013) suggest that an abnormally high ratio of dissolved inorganic nitrogen to dissolved reactive phosphorus in 2011 allowed the early summer cyanobacteria species *Aphanizomenon* to persist into July and August, coexisting with *Mycrocysits,* a species typical of later summer months. It is possible that cyanobacteria biomass in 2011 was driven primarily by nitrogen availability, a predictor not selected for the negative phase model by the GCV. Neglecting to distinguish between cyanobacteria species may limit the model's ability to capture the influence of community dynamics on overall cyanobacteria abundance throughout the summer.

While linear models are unable to entirely capture non-linear drivers of atmospheric and limnological processes, the NIPA approach highlights the diverse response of local and global predictor variables important to cyanobacteria productivity given the mean state of the atmospheric-oceanic system. Differences in the predictive power of phosphorus load and related variables (e.g. discharge, extreme events, suspended sediments) between MEI phases are particularly notable considering the large body of work establishing phosphorus load as a major driver of cyanobacteria productivity in Lake Mendota. The number of years in each phase is relatively small from a statistical perspective, however, compared to most inland lakes, the cyanobacteria biomass record for Lake Mendota is considered long. Nonetheless, continued collection of water quality data is warranted for the refinement of these models.

2.B.5 Conclusions

In this chapter skillful sub-seasonal cyanobacteria biomass prediction models are developed and compared with full-season prediction models to understand potential prediction gains and inform lake and beach management. The inter-annual variability of biomass results from a complex array of physical, chemical, and biological variables, many of which are significantly impacted by local climate, yet modulated broadly by large-scale climate phenomena through atmospheric teleconnections such as ENSO. In this chapter, spring and early summer variables are evaluated to determine their ability to represent within-season drivers of July-August cyanobacteria biomass.

In comparison to a traditional model conditioned on all years in the historical record, a two-phase approach is adopted – categorizing years as falling into either a positive or negative phase according to the pre-season MEI value. This modeling approach significantly improves predictions of July-August cyanobacteria biomass – particularly for *above normal* July-August conditions – and highlights the relative importance of unique local and global cyanobacteria biomass drivers in each phase. Notably, variables closely related to spring and summer phosphorus load are included in the negative phase model however are not significantly correlated with cyanobacteria biomass in the positive phase. This distinct behavior difference may be mediated by atmospheric teleconnections between ENSO and the Great Plains Low-Level Jet, which acts as a conduit for moisture transport from the mid-Atlantic to the Midwest. While inferences in how precipitation and thus variability in lake processes is modulated by ENSO specifically and large-scale climate generally are provided here, additional investigation is still warranted (Justić et al., 2005; Morse & Wollheim, 2014). Additional lines of inquiry could include development of coupled seasonal and sub-seasonal forecast systems for other water quality indicators, use of remote sensing methods to enhance observational records and predictability, and further integration of forecasts with lake and beach management alternatives.

Chapter 3. Large scale seasonal forecasting of peak season algae metrics in the Midwest and Northeast U.S.

Adapted from: Beal, M.R.W., Wilkinson, G. M., & Block, P. J. (2023). Large scale seasonal forecasting of peak season algae metrics in the Midwest and Northeast US. Water Research, 229, 119402. DOI: 10.1016/j.watres.2022.119402

3.1 Introduction

Rapid proliferation of algae in surface freshwaters has negative consequences for ecosystem function (Huisman et al., 2018; Sunda et al., 2006), economic opportunity (Dodds et al., 2009), and human health due to the potential for toxin production in some species (Carmichael, 2001; Carmichael & Boyer, 2016). In recent decades, anthropogenic disturbance of nitrogen and phosphorus cycles has resulted in widespread eutrophication, leading to an increase in the prevalence of harmful algae (O’Neil et al., 2012; Paerl & Paul, 2012; V. H. Smith, 2003). For many waterbodies, hydroclimatic variability plays an important role in determining water quality on inter- and intra-annual timescales, and may influence the suitability of conditions for algae growth (León-Muñoz et al., 2018; Scordo et al., 2022). Nutrient runoff, in particular, is sensitive to variability in the hydrologic cycle, which has been projected to intensify with climate change (Glavan et al., 2015; Me et al., 2018). Anthropogenic stressors favoring the dominance of harmful algae, combined with notable variability in algae biomass, presents a substantial challenge for water resource managers. In the U.S., harmful algae in large waterbodies such as Lake Erie has received significant research and media attention (Dalton, 2021; International Joint Commission, 2014; Patel & Parshina-Kottas, 2017; Reutter et al., 2011; Wines, 2014), however, despite similar concerns, strategies for managing harmful algae in small inland waterbodies across the U.S. have received less attention (Brooks et al., 2016).

In the northern hemisphere, algae biomass tends to peak in the late summer and early fall (July-October) as a result of a complex array of pre-season and within-season physical, chemical, and biological processes. In the Midwest and Northeast U.S., this season is characterized by warm temperatures and increased sunlight, allowing for increased photosynthesis and algae productivity (Singh & Singh, 2015). In many instances, significant intra- and inter-annual variability in peak season algae biomass is evident, driven partially by local hydrology and temperature that are in turn modulated by large scale climate phenomena through atmospheric teleconnections (Beal et al., 2021). Predictions of how these algae conditions vary may benefit lake managers by allowing them to take early actions to reduce or mitigate harm caused by intense algae growth. Short-term (days to weeks) predictions of chlorophyll-*a* (a proxy for algal biomass) are typically issued within-season and focus on expected bloom formation or toxin production, allowing managers to take rapid actions to address odor and taste issues, transition to alternative water sources, post warning signs at beaches, etc. (Zhang et al., 2013; Chen et al., 2015; Qian et al., 2021; Wynne et al., 2013). In contrast, longer-lead (months) pre-season predictions of expected algae conditions may allow lake managers to address a different set of actions (e.g. life-guard training, public awareness, etc.) and decisions, (e.g., testing and monitoring budgets and plans). Together, these predictions can provide decision makers with multi-scale information to inform appropriate actions at various lead times. However, season-ahead predictions for water quality have received relatively little attention.

Longer-lead predictions of oceanic chlorophyll-*a* and inland nutrient loading have been developed with some success (e.g. Cho et al., 2016; Park et al., 2019; Rousseaux et al., 2021), but little attention has been devoted to inland lakes. Long-lead predictions of algae that do exist typically focus on singular metrics (often mean biomass) to characterize the potential loss of ecosystem services due to algae accumulation, however, further characterization may be warranted. For example, Wilkinson et al., (2022) define three metrics to characterize algae conditions, including: *magnitude* (mean seasonal chlorophyll), *severity* (peak seasonal chlorophyll), and *duration* (length of time chlorophyll is above a threshold concentration). In addition to information provided by mean biomass, this approach provides lake managers with information that specifically addresses two key management concerns related to algae: the potential for severe consequences of algae blooms such as fish kills and toxin production, and length of time a lake may be unfit for recreation. Long-lead predictions also often rely predominantly on nutrient loads as predictors (R C Lathrop et al., 1998; C A Stow et al., 1997; Stumpf, Johnson, et al., 2016), however consideration of relevant hydroclimatic predictors has the potential to enhance prediction performance and expand the availability of water quality predictions to many small inland lakes (Beal et al., 2021). There is a large and long body of evidence illustrating the impacts of external nutrient loading on phytoplankton growth in lakes (J. A. Elliott et al., 2006; Kane et al., 2014; Reynolds, 1984; D W Schindler, 1978; Vollenweider, 1971). Phosphorus and nitrogen are widely considered the most important nutrients for phytoplankton growth in freshwater (D W Schindler, 1971; David W Schindler, 1977). Transportation of phosphorus and nitrogen into a lake from the surrounding watershed is an important driver of algae abundance in many systems. Nutrient transport is influenced by global and local hydroclimatic variables, and thus also represent important processes in determining algae abundance. Increased precipitation has been linked to increased fluxes of nitrogen and phosphorus (Sinha et al., 2017), particularly in extreme precipitation events (Stephen R Carpenter et al., 2015, 2018; Haygarth & Jarvis, 1997; Royer et al., 2006). Soil moisture conditions may also influence nutrient loading by regulating runoff potential (Kleinman et al., 2006; R. Liu et al., 2014). Finally, water temperature has also been shown to control phytoplankton biomass and growth rate (J. A. Elliott et al., 2006; Eppley, 1972; Konopka & Brock, 1978; X. Liu et al., 2019; Robarts & Zohary, 1987; Trombetta et al., 2021), and is closely linked to local air temperature (Shuter et al., 1983; Woolway et al., 2020; S. Zhu et al., 2020). March-June water temperature data are not readily available for all study lakes and are therefore not included in the final set of potential predictors.

In addition to management applications, season-ahead forecasts at scale provide a unique opportunity to understand ecological relationships between hydroclimatic variables and water quality (Houlahan et al., 2017). In particular, the relevance of global scale processes in determining algae biomass in inland lakes is not well studied. Several studies have identified teleconnections between large-scale climate phenomena and phytoplankton dynamics in inland lakes (Arhonditsis et al., 2004; da Rosa Wieliczko et al., 2021; Xiao et al., 2019), however, few studies exist that investigate the application of global climate patterns to chlorophyll-*a* prediction in inland lakes. Beal et al. (2021), developed a sub-seasonal (2-month lead) forecast of cyanobacteria biomass in Lake Mendota, Wisconsin (WI), conditioned on local hydroclimatic variables and teleconnections with global climate patterns. A large-scale analysis of season-ahead predictors of algae biomass is well suited to improve the understanding of dominant climate signals related to chlorophyll-*a*.

Using chlorophyll-a time series from 178 lakes in the Northeast and Midwest U.S. we evaluated if global and local hydroclimatic processes can be used to predict algal magnitude, severity, and duration in each lake using a statistical modeling and forecast validation (hindcast) approach.

Specifically, we address the following questions:

1) Do local and global (sea surface temperature, SST) hydroclimatic variables correlate with chlorophyll metrics in a given lake?

2) Are skillful predictions based on these variables possible for algal magnitude, duration, and severity?

3) Can variability in forecast model performance be explained by static, lake-specific characteristics?

This modeling approach may provide insight into the role of local and global hydroclimatic variability in the development of peak season algal biomass, evaluate the ability of hydroclimatic variables to provide actionable information to lake managers at a seasonal timescale, and indicate which characteristics of small inland lakes make them ideal candidates for seasonal forecast development.

3.2 Materials and Methods

3.2.1 Lake Characterization and Selection

Chlorophyll-*a* measurements, sampled at the surface of the lake and analyzed following project-specific protocols for each lake, were obtained from LAGOS-NE (Patricia A. Soranno et al., 2017). The measurements range from 1982 to 2013 in the database, but several chlorophyll timeseries were extended through 2020 by collating additional measurements from the reporting agency or program referenced in LAGOS-NE for each lake. LAGOS-NE aggregates data from lakes located throughout the Midwest and Northeast U.S. In this region, chlorophyll-*a* tends to reach peak concentrations between July and October (JASO) (Figure 3-1). The data from this period were used to calculate three chlorophyll metrics: *magnitude, severity*, and *duration* (see explanation below) for each lake year. To adequately characterize July-October chlorophyll-*a* metrics, sufficiently long observational records and frequent within-season sampling are needed. Therefore, lakes for this analysis needed at least 15 years of July-October chlorophyll-*a* measurements and a minimum sampling frequency of once every 14 days, following the selection methods of Wilkinson et al. (2022). Based on these requirements, 178 lakes were identified from 10 states in the Midwest (Michigan, Wisconsin, Minnesota, Ohio, and Missouri; 64 lakes) and Northeast (New York, Vermont, Rhode Island, Pennsylvania, and Maine; 114 lakes) U.S. The chlorophyll data were log-transformed for analysis to create a Gaussian-like distribution. Selected lakes had an average depth of 15.8 meters (min=1.2 m, max=198.4 m) and an average area of 1297.9 hectares (min=1.4 ha, max=113,496.5 ha).

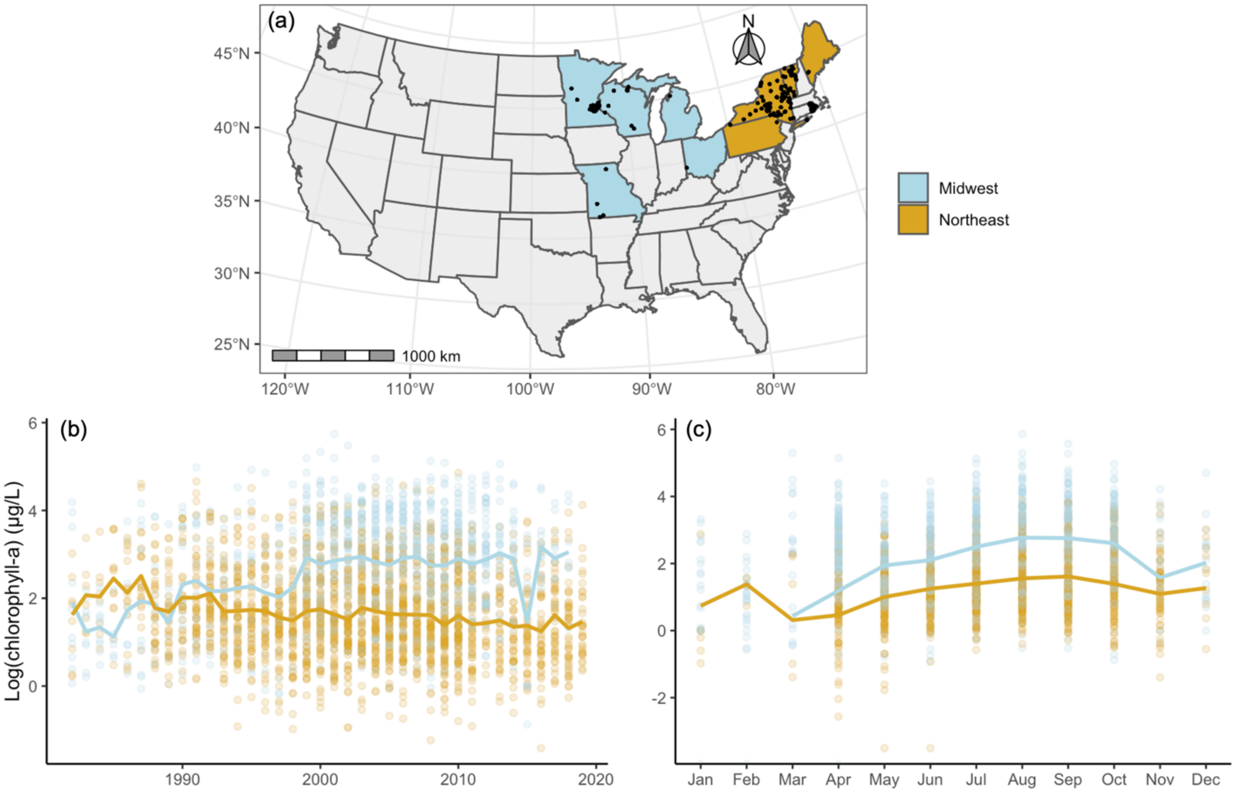

Figure 3-1. Candidate lakes for model development in the Northeast and Midwest U.S. (a), including mean JASO chlorophyll-a (μg/L) concentrations (b) and mean monthly chlorophyll-a values (c). Points in (b) and (c) represent values for each lake.

3.2.2 Chlorophyll Metrics

To characterize (and eventually predict) algal conditions in each lake year during the July-October season, three metrics were used: *magnitude*, *severity*, and *duration* of chlorophyll-*a,* as defined by Wilkinson et al. (2022). *Magnitude* is the mean chlorophyll-*a* concentration in each lake year, and *duration* is the portion of the season during which chlorophyll-*a* concentrations exceed a threshold concentration for recreational value based on Angradi et al. (2018). Here, the *severity* metric has been altered from Wilkinson et al. (2022) and is defined as a function of *magnitude*. Seasons in which *magnitude* exceeds the 95th percentile of all historical chlorophyll concentrations (rather than year specific concentrations) are categorized as *severe*. On an inter-annual timescale, *magnitude* characterizes the average conditions during the peak season and the corresponding impacts on ecosystem services. *Severity* reflects the probability of extreme algae biomass (*magnitude)* that is most likely to result in severe consequences like toxin production and fish kills. Finally, *duration* is the persistence of high algal biomass associated with a loss of recreational value during the summer, peak season. The threshold concentrations used here are developed for two ecoregions (“Mountains” and “Plains”) and vary in each region based on recreational user’s expectations for water quality. Because these are large regions, lakes that fall below or above the chlorophyll-a threshold in nearly all sampling events are removed from analysis to avoid artificially inflating overall forecast skill. Together, these metrics may provide an enhanced understanding of algae conditions during the peak season of algal production, with prospects for more refined actionable information. Compared to a singular forecast of mean chlorophyll (magnitude), a forecast that additionally provides advanced warning of protracted water quality impairment (duration) and the potential for severe consequences of algae growth (severity) allows for more nuanced decision making around budgeting, testing, and communicating water quality expectations with the public. The extent to which these three metrics are correlated with local and global climate variables and predictable across a diverse set of lakes is the focus of this work.

3.2.3 Predictor variable selection

To address whether local and global hydroclimatic variables are correlated with chlorophyll metrics in a given lake, we evaluate to what extent pre-season (March through June) observations of local and global hydroclimate variables are correlated with magnitude and duration for each lake. Correlations for *severity* were not evaluated independently in this analysis as the *severity* metric is a function of *magnitude*. In addition to identifying common local and global hydroclimatic variables correlated with chlorophyll metrics, this analysis was used to identify variables that would be used in the development of each lake-specific forecasting model, and the validation of each model in a hindcasting analysis. The hindcasting analysis uses the lake-specific forecasting model to predict each year in the chlorophyll metric timeseries without predictor information from the year of interest, simulating a forecast for model validation. Specifically, we included March-June variables from readily available, gridded datasets that were connected to physical processes that may affect *magnitude* and *duration* including (Table 3-1): total precipitation (mm), mean air temperature (°C), mean volumetric soil moisture (m^{-3}/m^{-3}), the sum of daily precipitation events exceeding 20 (40) mm for Midwest (Northeast) watersheds, and global sea surface temperature (SST) anomalies (°C). Methods for evaluating SST anomalies as predictors are described below. Additionally, we evaluated if pre-season chlorophyll-a, which reflects in-situ processes and the nutrient availability at the start of the season, is correlated with peak-season chlorophyll metrics. Excluding extreme events, all metrics were averaged over the March-June season.

As discussed previously, local hydrology regulates nutrient transport into lakes from the surrounding watershed, which may ultimately influence algae abundance. The influence of local hydrology-based candidate predictors (precipitation, extreme events, and soil moisture) may vary based on land use and topography within a lake's watershed. Therefore, local hydrology predictors were evaluated using a correlation analysis at each grid within each lake's HUC12 watershed. HUC12 watershed polygons (Watershed Boundary Dataset, 2021) were subset to only include areas higher in elevation than the corresponding lake (Figure 3-A.1) using gridded elevation data for each watershed from the elevatr package for the R statistical programming language (Hollister et al., 2021). The timeseries of candidate predictor variables from each grid intersecting the watershed polygon was used in the correlation analysis as was an average of all intersecting grids. The grid with the strongest, statistically significant correlation was retained for the subsequent hindcasting analysis. High precipitation events may have a more significant influence on overall nutrient loading than total precipitation (Stephen R Carpenter et al., 2015), however, precipitation events that lead to large loading events may vary by region due to land use topography, and nutrient availability. Therefore, separate thresholds were chosen for extreme precipitation events in Midwest (20 mm) and Northeast (40 mm) watersheds based on a sensitivity analysis of significant correlations between high precipitation events and peak-season chlorophyll-*a magnitude* conducted for each region.

Table 3-1 Variables used in correlation analysis with peak season chlorophyll-*a* metrics, including data source and resolution.

Predictors (March-June)	**Source**	**Resolution**
Total precipitation (mm)	PRISM	4km
Mean air temperature (°C)	PRISM	4km
Mean volumetric soil moisture (m^{-3}/m^{-3})	Copernicus Climate Change Service	0.25°
Precipitation events exceeding 20 (40) mm in Midwest (Northeast) watersheds	PRISM	4km
Sea surface temperature anomalies (°C)	NOAA ERSST	2°
Pre-season Chlorophyll-*a* (μg/L)	LAGOS	In-situ

On a global scale, pre-season sea surface temperatures can influence in-season precipitation and temperature over the U.S. through modulation of atmospheric flow and thus indirectly influence peak-season chlorophyll metrics (Barnston, 1994; Giannini et al., 2000; Markowski & North, 2003). SSTs evolve slowly, with persistent (months to years) anomalies, and thus can serve well as predictors at seasonal timescales (Barnett, 1981). To identify oceanic regions with strong teleconnections to the Northeast and Midwest U.S., global pre-season sea surface temperature (SST) anomaly grids were correlated with each lake's magnitude timeseries (Figure A.2). Not surprisingly, correlation patterns, and thus oceanic regions of influence, vary between the Midwest and Northeast U.S. (Ropelewski and Halpert, 1987; CPC, 1997; Enfield et al., 2001; Tootle et al., 2005), therefore identification of teleconnections is performed separately for the Midwest and Northeast. The number of significant correlations with chlorophyll-a timeseries were tallied for each SST grid and mapped to identify oceanic regions in which SST grids were associated with algae abundance in the Northeast and Midwest. SST grids that were significantly correlated with a large fraction of lakes (≥10 for Midwest, ≥20 for Northeast) were applied to a principal component analysis (PCA) to extract the dominant modes of variability in SST data and reduce the dimensionality of candidate SST predictors. Given the large number of SST grids, there is a high likelihood of generating spurious correlations. Performing PCA extracts the dominant climate signals and minimizes the effect of spuriously correlated grids. Principal components (PCs) that explained more than 5% of the variance in SST anomaly data were retained as candidate predictors.

The time series of variables in Table 1 were used in a correlation analysis with chlorophyll *magnitude* and *duration* for each lake to identify predictor variables for forecast model development and the subsequent hindcasting analysis (Figure 3-2). For each lake, all variables that were significantly correlated with the chlorophyll metric ($p<0.05$) were retained for the forecasting model. Out of the 178 lakes evaluated, 135 lakes (50 Midwest; 85 Northeast) had at least one significant predictor variable for magnitude and 82 lakes (30 Midwest; 52 Northeast) for duration. *Severity* is a function of *magnitude* and therefore retained the same set of predictors.

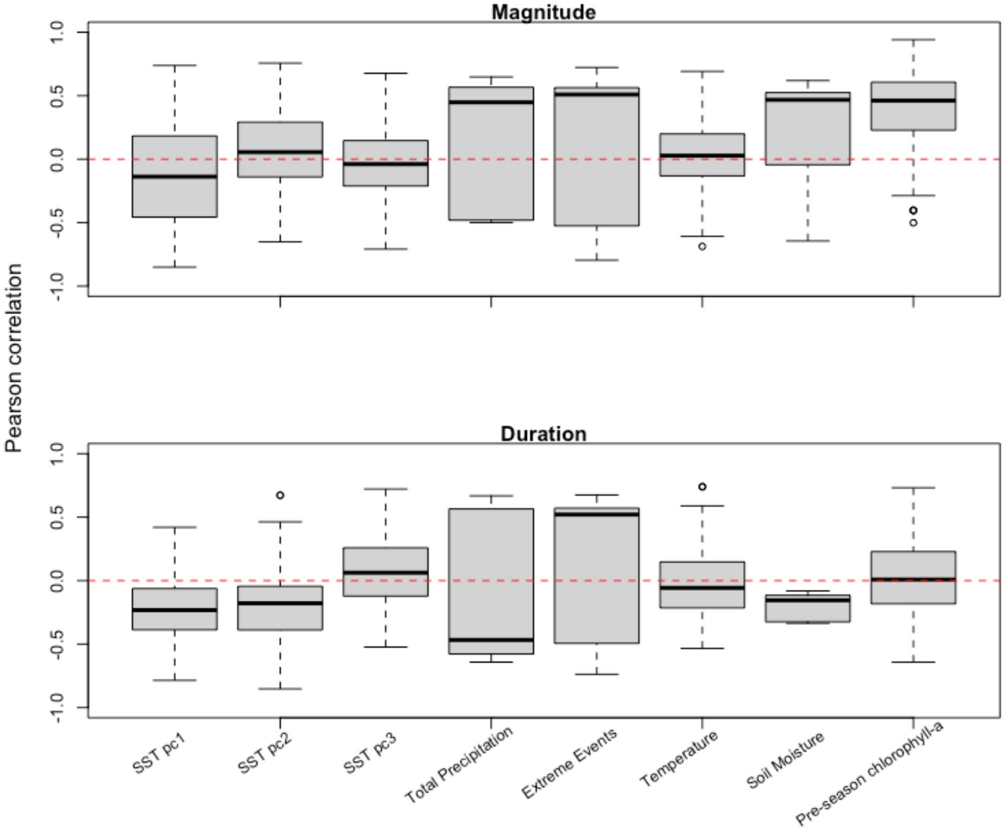

Figure 3-2. The distribution of Pearson correlations between candidate predictors and algae metrics for all lakes.

3.2.4 Forecast model development

Two forecasting models were developed for each lake, one focused on *magnitude (including severity*) and a separate model for *duration* (proportion of sampling events above the impairment threshold). The array of processes and feedbacks influencing algae growth and abundance are notoriously complex (Glibert & Burkholder, 2006; Ho et al., 2019; Roelke & Buyukates, 2001), motivating a statistical modeling approach over a process based/physical model approach. For lakes with only one significant predictor from the correlation analysis, a simple linear regression between that variable and the chlorophyll metric was constructed for the model to be used in the hindcast analysis. For lakes with multiple significant predictors, a principal component analysis and regression approach was used to build the forecasting model. PCA effectively deals with any multi-collinearity present between predictors and therefore does not artificially inflate predictive skill. Here, principal components were retained for the forecast model if they explained more than 10% of the variance. This modeling approach assumes relationships between candidate predictors and peak season algae metrics on a seasonal timescale to be linear, however, given that many drivers of algae growth on short timescales (days to weeks) are considered nonlinear processes, model residuals were evaluated for evidence of nonlinear relationships. Autocorrelation was also investigated in each of the candidate predictors and algae metrics. Except for SST pc1, which likely captures baseline increases in the temperature of the Pacific Ocean, less than 10% of timeseries for each variable had more than two statistically significant autocorrelations (lag 1-10). Additionally, random forest regression, a nonlinear, nonparametric modeling approach, was tested to determine if there were notable changes in model skill due to potential nonlinearities or autocorrelation. Statistical models were developed using R version 4.2.1.

A leave one out cross-validation approach was used to evaluate model performance for each lake and chlorophyll metric (hindcasting). In short, for each lake the observed value from one year of the chlorophyll metrics was removed from the timeseries and the forecasting models (above) were used to predict the missing value. This process was done iteratively for all years in the timeseries for all lakes individually and both *magnitude* (including *severity*) and *duration*. A prediction ensemble was created for each peak-season chlorophyll metric in a lake year based on model errors (difference between observed and predicted chlorophyll-a) across the hindcast at that lake. Ensemble members are generated from a normal distribution of errors with mean zero, based on maximum likelihood estimation. For each time-step, 100 random draws from the distribution are added to the *magnitude* and *duration* predictions to form the ensemble prediction (Alexander et al., 2019; Helsel & Hirsch, 1992).

3.2.5 Performance Measures

To assess model performance, four measures were adopted: correlation coefficient (R^2), root mean square error (RMSE), ranked probability skill score (RPSS), and Heidke skill score (HSS) (Epstein, 1969; Heidke, 1926). HSS and RPSS are measures of categorical skill, interpreted as a percent improvement over a reference forecast. A standard forecast for hydro-climate prediction is an equal-odds (climatological) distribution of historical observations. Here, the distribution of historical observations of chlorophyll-a metrics for each lake are split into four categories representing *below normal, near normal, above normal,* and *severe* algae conditions. If no predictive information is present, a probabilistic prediction of JASO chlorophyll-a would default to climatology (33% chance of *below normal*, 33% chance of *near normal* conditions, 28% chance of *above normal*, and 5% chance of *severe* conditions). Similarly, observations of duration are split into two categories based on mean duration to represent expected below and above normal conditions. Forecast models developed for magnitude and duration generate probabilistic predictions of each category that are compared against climatology (equal odds). This allows for a direct comparison between the forecast models developed here and a benchmark climatology model to understand the prospects for enhanced predictive skill. In general, prediction models outperform climatology when the predicted probability of the observed category is greater than the climatological probability (e.g. 50% for two categories, 33% for three categories). HSS is defined in previous chapters. HSS values range from -∞ to 1, where negative values represent a forecast that performs worse than climatology, 0 represents no skill, and 1 represents a perfect prediction model. The RPSS, defined previously, is a categorical skill score that increasingly penalizes an ensemble forecast for assigning greater probability to categories farther from the observed category. RPSS values range from -∞ to 1, where 0 represents no skill and 1 represents a perfect forecast. RPSS values are calculated for each year and the median value is reported.

Finally, we evaluated if variability in hindcast model performance (forecasting skill) among lakes was related to static characteristics of the ecosystems. We compared forecasting skill among categories of trophic state, lake area, land cover, and geographic region among lakes. The trophic state index (TSI) is calculated based on chlorophyll-*a* and is categorized as oligotrophic (TSI<40), mesotrophic (40≤TSI<50), eutrophic (50≤TSI<70), and hypereutrophic (TSI>70) (Eq. 5) (Carlson, 1977).

$$\text{(6)} \qquad \mathrm{TSI(CHL)} = 9.81 \ln(\mathrm{CHL}) + 30.6$$

3.3 Results

3.3.5 Leading Algae Characteristic Predictors

The three most frequently retained predictors include pre-season chlorophyll-*a*, PC1 from SSTs, and extreme events for *magnitude* models. The most frequently retained predictors for *duration* models include pre-season chlorophyll-*a*, and PC1 and PC2 based on SSTs. PCs derived from SSTs typically represent dominant large-scale climate signals, potential physical processes are explored further in the discussion. *Magnitude* and *duration* metrics are uncorrelated in most lakes (only 11are significantly correlated at the 95% confidence level), suggesting unique seasonal drivers for each metric in most lakes. Compared to *magnitude* models, *duration* models have a more even distribution of retained predictors (Figure 3-3). In *magnitude* models pre-season chlorophyll-*a* meets the selection criteria in 63% of models, SST PC1 in 42% of models, and extreme events in 16%. In *duration* models, SST PC2 is selected in 38% of models, SST PC1 in 33%, and pre-season chlorophyll-*a* in 20%. In both *magnitude* and *duration* models, all three predictors are unlikely to appear in the same model, suggesting variable influence of these processes by lake. In *magnitude* models, all three of the most frequently retained predictors are included in 7% of lakes. In contrast, one of the three predictors in included in 87% of *magnitude* models. In *duration* models, all three of the most frequently retained predictors are included for only two lakes (2.5%), while one of the three predictors is included in 71% of *duration* models.

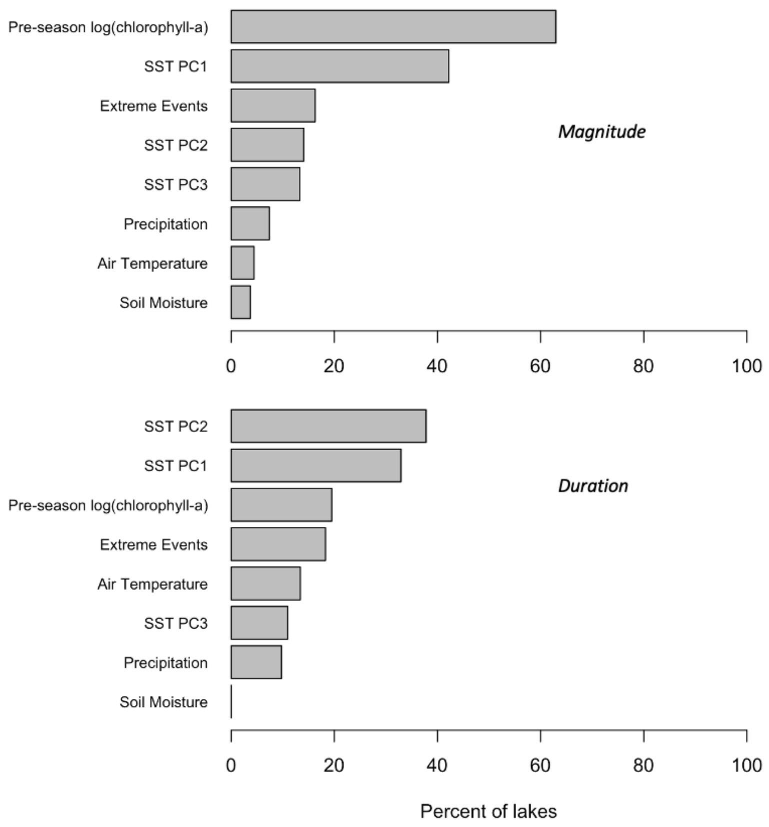

Figure 3-3. The percent of lakes for which each predictor is statistically significantly correlated with *magnitude* (top) and *duration* (bottom); 178 lakes evaluated.

Preseason SST grids that are statistically significantly correlated with chlorophyll-*a* timeseries from at least 20 (10) lakes are retained for the Northeast (Midwest) region (Figure 3-4). SST regions retained for both Midwest and Northeast lakes indicate teleconnection signals from the northern Atlantic and the equatorial pacific. For Midwestern lakes, more grids are retained in the mid and northern Atlantic compared to the Pacific. Comparatively, the strongest signals for northeastern lakes are split more evenly between the upper Atlantic and the equatorial pacific.

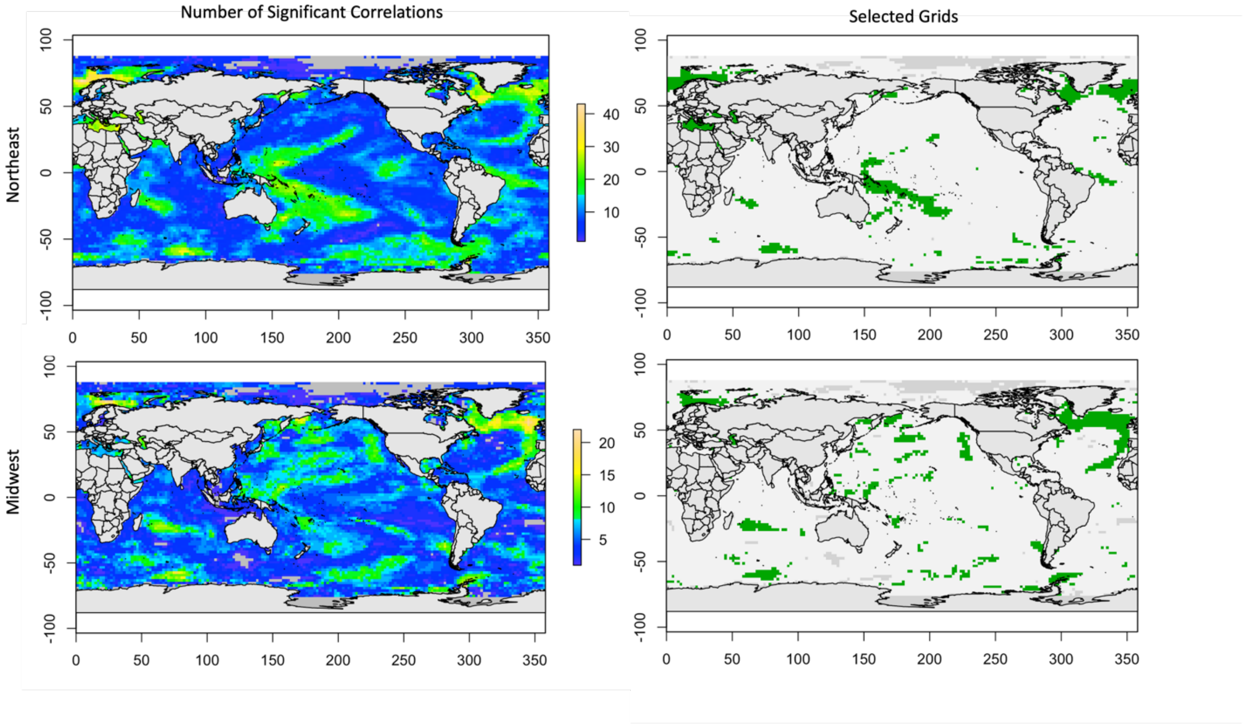

Figure 3-4. Number of preseason SST grids that are statistically significantly correlated with chlorophyll-*a* timeseries (left column) from lakes in the Northeast (top row) and Midwest (bottom row). Grids retained (right column) have at least 20 (10) significantly correlated lakes for the Northeast (Midwest).

While local hydroclimatic predictors are present in a significant proportion of *magnitude* and *duration* models, in most instances they are retained with SSTs or pre-season chlorophyll-*a,* particularly in *magnitude* models. Some combination of local hydroclimatic predictors (precipitation, extreme events, soil moisture, and air temperature) are retained in 27% of *magnitude* prediction models, while just 9% *of magnitude* models utilize only local hydroclimatic predictors (i.e., without SST PCs or preseason chlorophyll-*a)*. Compared to *magnitude* models, a larger proportion of *duration* models are built solely with local hydroclimatic variables. Hydroclimatic predictors are retained in 38% *duration* models, and 22% of models are built exclusively with hydroclimatic predictors (Figure 3-A.3). Finally, 86% of lakes have 1 or 2 significant predictors while the maximum number of predictors retained for a lake is 5 (Figure 3-A.4).

3.3.6 Model Performance

For *magnitude* prediction models, simple linear regression (single predictor) is applied to 53% of lakes, whereas principal component regression (PCR, principal component analysis with multiple linear repression) is applied to 47% of lakes. A cross-validated hindcast assessment for all lakes results in a mean R^2 value of 0.28 (0-0.85) and a mean RMSE of 0.47 (0.29-1.21). For categorical performance, the mean HSS and RPSS values are 0.17 and 0.10, respectively. Additionally, 87% (70%) of lake models have HSS (RPSS) values greater than zero, indicating an improvement of prediction skill over climatology for most lakes. Further, these models predict *above normal* and *below normal* algae abundance moderately well (Figure 3-5), with RPSS values of 0.39 and 0.30 respectively. *Severe* events prove difficult to predict; approximately 51% of *magnitude* models accurately predict an *above normal* or *severe* year when an elevated algae event is observed (i.e., not *below* or *near normal*) for more than half of observed elevated algae events (Figure 3-6).

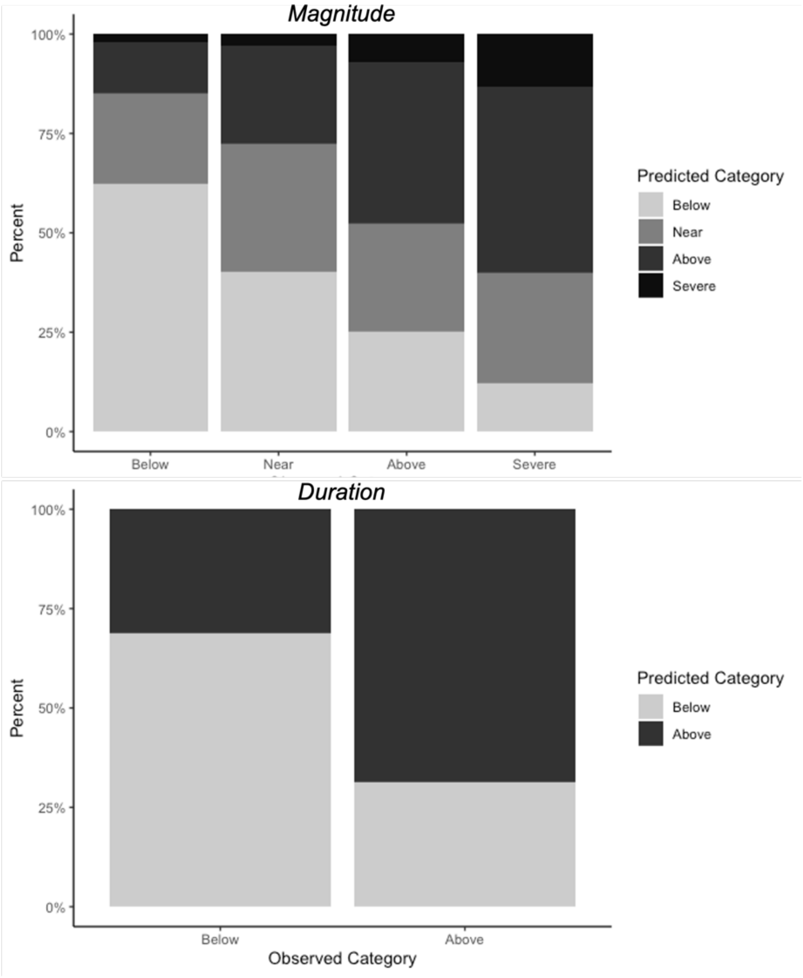

Figure 3-5. The proportion of each prediction category for each observed category (all lakes).

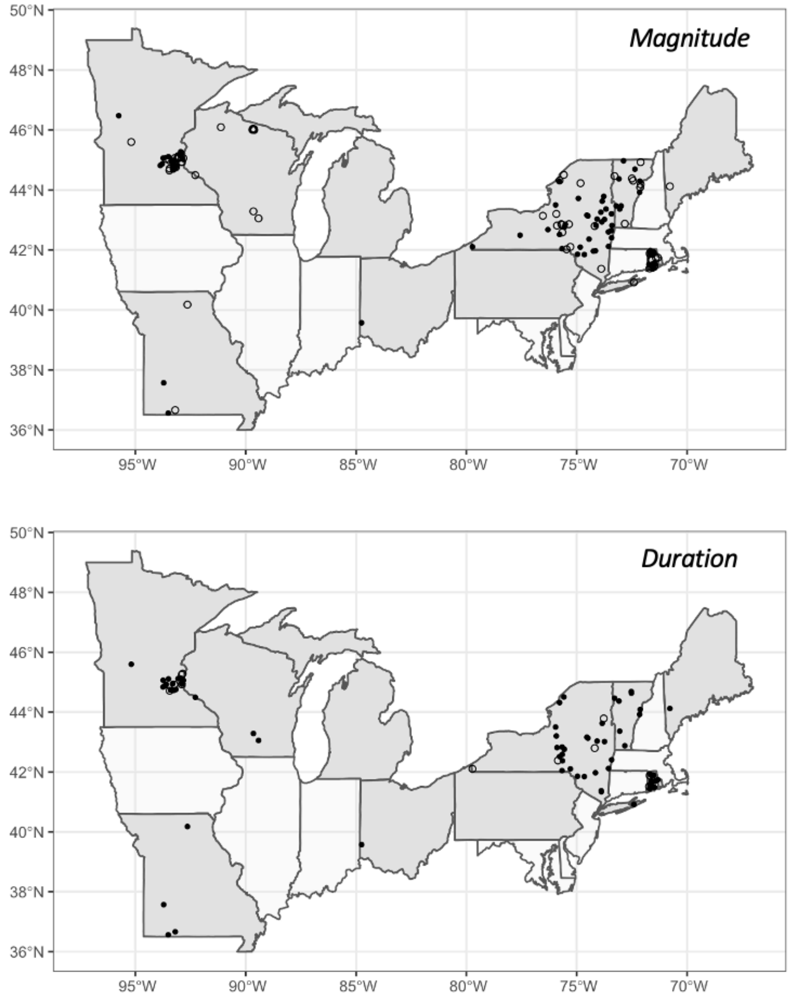

Figure 3-6. Lake models that correctly predict more (less) than half of elevated *magnitude* (top) and *duration* (bottom) events, illustrated as open (closed) circles.

For *duration* prediction models, simple linear regression is applied to 78% of lakes and PCR to 22% of lakes. Mean R^2 and RMSE are 0.23 (0-0.65) and 0.34 (0.16-0.42), respectively; mean HSS and RPSS scores are 0.39 and 0.42, respectively, and 96% (94%) of models improve over climatology based on HSS (RPSS). In a hindcast assessment, *durations* of *above normal* are correctly predicted for more than half the available timeseries in 87% of lakes (Figure 6). Additionally, considering all models, 67% of *above normal durations* are accurately predicted (Figure 3-5), a stark improvement over climatology for most lakes.

Average model skill is similar between regions is similar, however, Northeast lake forecast models outperform Midwest lake models where differences in average skill scores occur (Figure 3-7). This may be expected given that the primary difference in predictor selection between models in the Northeast and Midwest is selection of relevant SST grids. Relevant Northeast SST regions are more coherent than regions for the Midwest, which may represent a stronger influence on lake processes. The minimal differences may also point to consistency in the predictive power of local and within-lake variables between the Northeast and Midwest.

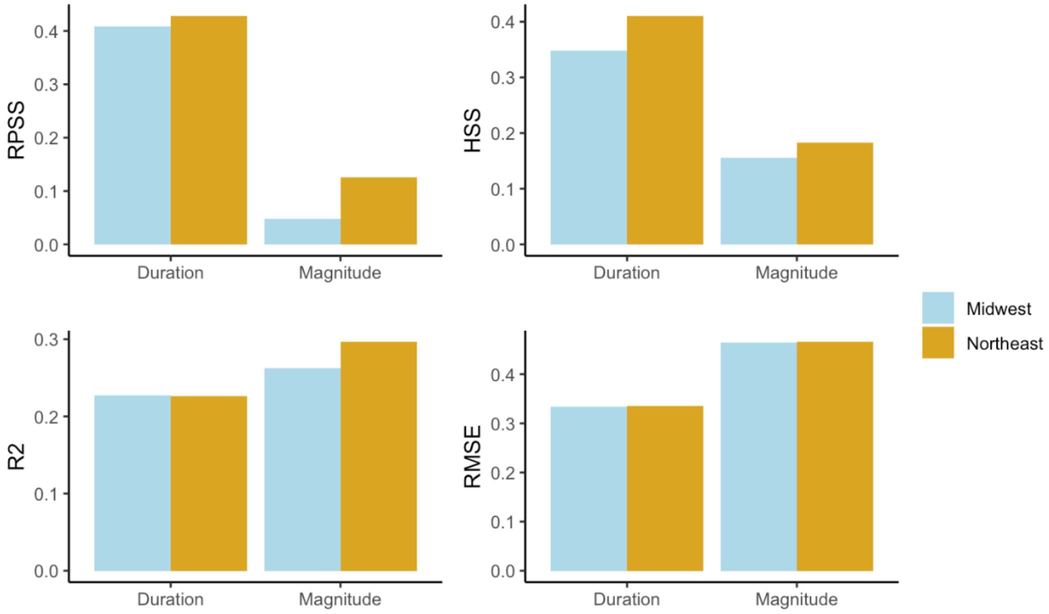

Figure 3-7. Average model skill scores for *duration* and *magnitude* models by region.

To evaluate the presence of nonlinear relationships, residuals between predicted and observed chlorophyll metrics were investigated for each lake. Residuals generally appeared random, suggesting that the relationships between the pre-season drivers and peak season chlorophyll metrics investigated for this analysis can be approximated as linear. Hindcasts were also generated using random forest regression to test for an increase in predictive skill, which may indicate the presence on nonlinear relationships. Random forests models were created with the same predictors selected for PCR, each with 500 trees. Cross validated hindcast results are similar or slightly worse than PCR for both *Magnitude* (Mean: $R^2 = 0.25$, RMSE = 0.48, HSS = 0.12, RPSS = 0.04) and *Duration* (Mean: $R^2 = 0.23$, RMSE = 0.34, HSS = 0.34, RPSS = 0.39), suggesting that PCR is a suitable approach.

3.4 Discussion

3.4.1 Regional Characteristics

Predictions of peak season algae growth at scale provide insights into relevant global and local-scale processes setting the conditions for peak season algae biomass. Pre-season SSTs and chlorophyll-*a* observations provide the most predictive power for both *magnitude* and *duration* metrics, reflecting the importance of these scales. However, while SSTs and chlorophyll-*a* are selected most frequently as predictors for both *magnitude* and *duration*, the importance of each predictor is mixed by region (Figure 8). Given that SST-atmosphere teleconnections typically have regional influence, spatial variability in performance of SST predictors may be the result of localized pre-season processes superseding the influence of large-scale climatic processes in peak season algae biomass. For example, food web dynamics are well established as having significant influence on aquatic primary productivity (S R Carpenter et al., 1987; Lampert et al., 1986; Vanni & Temte, 1990). In this study, however, the representation of food web dynamics is limited to pre-season algae abundance. While this variable is shown to be a powerful predictor of peak season productivity in many lakes, the effect of pre-season predatory control on algae communities is unrepresented. This may limit the skill of prediction models for lakes in which zooplankton grazing plays a significant role in determining algae populations in the summer and early fall. This limitation may be responsible for differences in average model performance. For example, *magnitude* and *duration* models in which pre-season chlorophyll-*a* is an important predictor, and SSTs are not, perform worse on average than the inverse (Table 2). Lakes retaining only pre-season chlorophyll-*a* may be more dependent on within-lake processes, many of which are not represented in these models. This may explain the less robust predictive signal from pre-season chlorophyll-*a* in these lakes, compared to lakes more heavily influenced by SST-atmosphere teleconnections.

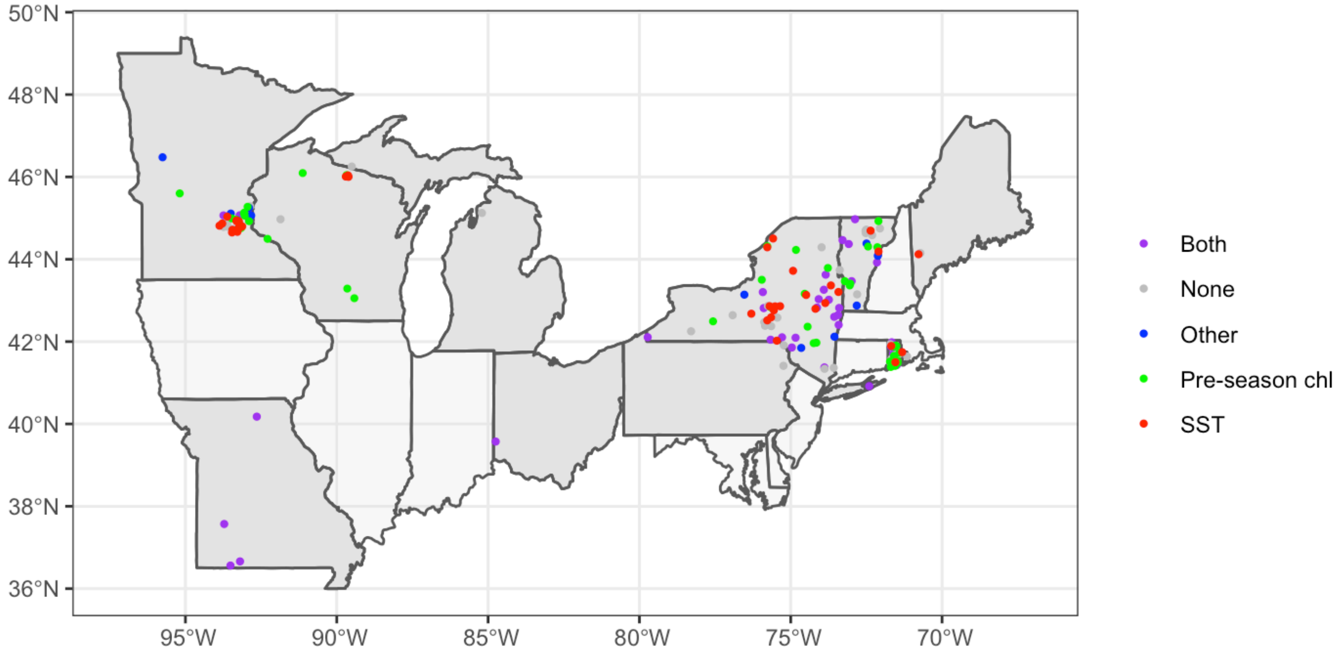

Figure 3-8. Lakes in which chlorophyll-*a*, SSTs, both, or neither are selected for *magnitude* predictions.

Table 3-2 Average skill scores for *magnitude* and *duration* models in which pre-season chlorophyll-*a* and/or SSTs are utilized as predictors.

Variable	High/Severe Correct	RPSS	HSS	R^2	RMSE	Predictand
Both	0.58	0.22	0.20	0.43	0.49	*Magnitude*
MAMJ chlorophyll-*a*	0.47	0.03	0.15	0.24	0.50	
SSTs	0.51	0.08	0.16	0.24	0.40	
Other	0.44	0.08	0.14	0.17	0.43	
Both	0.65	0.55	0.40	0.31	0.33	*Duration*
MAMJ chlorophyll-*a*	0.56	0.12	0.10	0.18	0.36	
SSTs	0.70	0.43	0.41	0.21	0.34	
Other	0.67	0.22	0.27	0.11	0.33	

As discussed previously, relevant SST anomaly grids are identified for the Northeast and Midwest separately. The PCs of selected SST anomaly grids represent the dominant climate signals affecting the selected lakes across the Northeast and Midwest U.S. and may therefore be associated with large-scale climate phenomena that have well-established teleconnections with climate conditions across North America. Two dominant climate phenomena with variable impacts on local hydroclimatic conditions across the Northeast and Midwest include the North Atlantic Oscillation (NAO) and the El Niño Southern Oscillation (ENSO) (Ropelewski & Halpert, 1987; Visbeck et al., 2001). In the Northeast, two of the three SST PCs used as potential predictors are significantly correlated with both the NAO index and multivariate ENSO index (MEI). In the Midwest, two PCs are significantly correlated with the MEI and all three are significantly correlated with the NAO index. This suggests that interannual variability in local climate, and in lakes, across the Northeast and Midwest is associated with both ENSO and NAO-like climate signals (Figure A.5).

Compared to pre-season chlorophyll and SSTs, other variables considered here provide only modest skill in predicting algae characteristics for most lakes. Despite the perceived importance of local land and hydrologic variables modulating inflow and lake processes, few were retained as predictors; however, land use and other watershed characteristics may be important in determining the relevance of these predictors. Local hydrology might be expected to play a larger role in promoting algae growth in agricultural watersheds, given the effect of runoff on nutrient loading (Castillo et al., 2000; Mander et al., 2000). This is reflected in predictor selection for model construction, for example, *magnitude* models in watersheds with greater than 25% agricultural land are nearly twice as likely to retain a local hydrologic predictor compared to models in watersheds with less agricultural land (18.4% vs. 36%). March-June air temperatures are also retained in relatively few forecasting models overall but are notably included in more *duration* models than any of the local hydrologic predictors. As discussed previously, the importance of temperature in determining algae growth is well established, however, the air temperature predictor included here may be too simplistic to capture the relationship across all lakes. Significant variability in prediction skill exists among lakes with little evidence of spatial patterns. On average, skill scores are higher in the Northeast compared to the Midwest, however, variability within both regions is significant. The frequency of predictor retention and the average magnitude of significant correlations between predictor variables and algae *magnitude* are similar by region particularly for the most frequently selected predictors, including SSTs, pre-season chlorophyll-a, and extreme events (within 0.05). SST PC1 and extreme events in the Northeast have slightly higher correlations with algae *magnitude* than in the Midwest, which may help explain slightly higher model skill scores in the Northeast. The magnitude of correlation between predictor variables and algae *duration* is also similar, however, SST PC2 is retained much more often in Northeast models compared to Midwest models (25% of lakes vs 5% of lakes). Given that SST PC2 is correlated with the NAO, this might indicate a greater influence of the NAO on *duration* in the Northeast and help explain the slightly higher skill scores of *duration* models in the Northeast.

Similarly, distributions of skill scores across trophic state and lake area are variable (Figure 9). On average, magnitude models appear to accurately predict increased algae abundances more frequently in larger lakes and in mesotrophic and oligotrophic lakes. Duration models have approximately equal distributions of skill across lake area and trophic state. While the differences in magnitude model skill based on lake area and TSI category are notable, they were found to be statistically insignificant in an analysis of variance (ANOVA; lake area P =0.24, TSI P = 0.36), therefore it may be difficult to draw definitive conclusions from these results. There are a few potential explanations for the variability in magnitude model performance. In an analysis of lake size and primary productivity in the Canadian Shield lakes, Fee et al., (1994) found that larger lakes more efficiently convert external nutrient loads into phytoplankton biomass due to more frequent resuspension of sediments, and receive a higher proportion of nutrient loads from runoff rather than direct precipitation. The focus in this analysis on hydroclimatic predictors of nutrient runoff is therefore consistent with increased predictive skill in larger lakes. Notably, air temperature did not stand out as an important predictor of algae abundance at a season-ahead lead time. As discussed previously, temperature is well-established as an important variable in algae growth. In some cases warm temperatures have been shown to hold greater influence over phytoplankton growth (Salmaso et al., 2012), and cyanobacteria growth in eutrophic lakes (Rigosi et al., 2014). The air temperature predictor used here may be too simplistic to capture peak season water temperatures, which may disproportionately contribute to lower model performance in eutrophic lakes. Rusak et al., (2018), found a positive relationship between variability of chlorophyll-a and trophic status in 18 globally distributed lakes, which may also reduce predictability (Cottingham et al., 2000). This may explain the moderate reduction in average skill in magnitude models for eutrophic lakes compared to lakes of a lower trophic status.

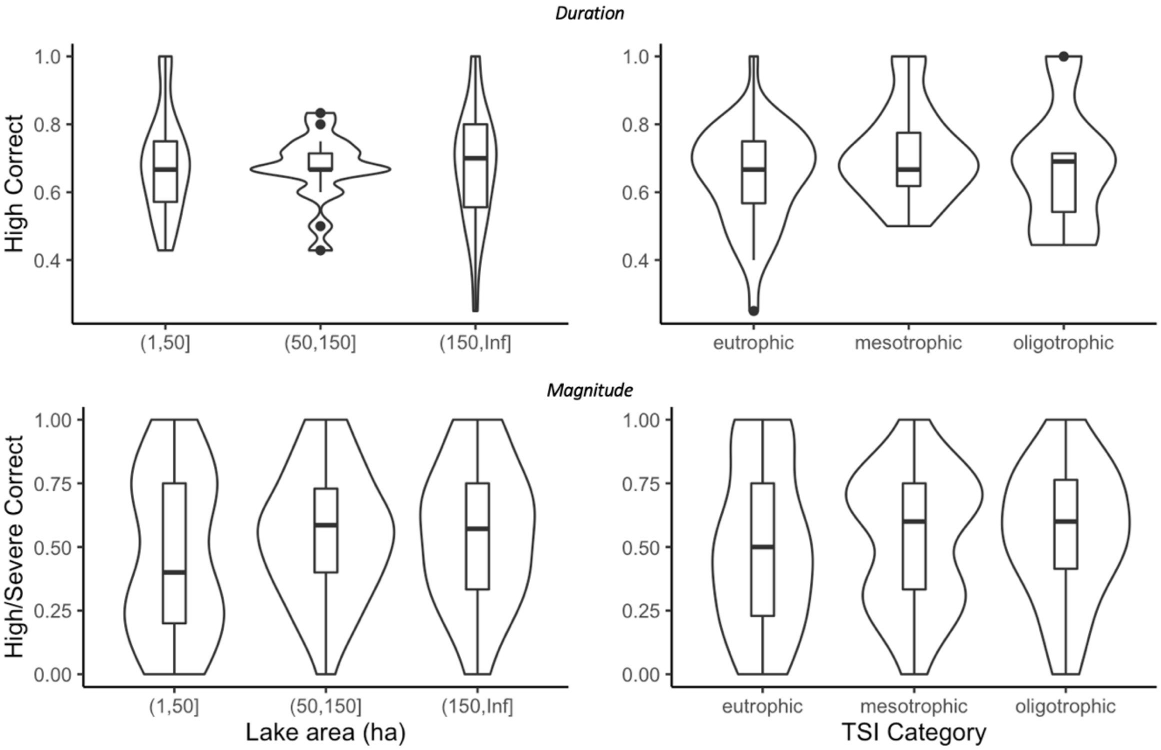

Figure 3-9. Violin plots of the proportion of *above normal* or *severe* events correctly predicted for each algae metric, compared to lake area (ha) and Trophic State Index (TSI).

Variability in prediction skill can also exist between similar or proximal lakes. Mariaville lake and Duansespurg reservoir are two eutrophic waterbodies located in eastern New York, approximately five miles apart. Mariaville lake has one of the best performing *magnitude* prediction models among the lakes considered and accurately predicts *above normal* and *severe* chlorophyll-*a* conditions for each of the five years in which they are observed. Comparatively, Duansespurg reservoir only correctly predicts two out of six observed events (Figure 3-10). In 1999, for example, the Mariaville lake model accurately predicts a large probability of *above normal* conditions (observed state) and is even able to differentiate between *above normal* and *severe*. Comparatively, for the same year, the Duansepurg model only predicts a 1% chance of *below normal* conditions (observed state), and both models predict an approximately 80% chance of *above normal* or *severe* conditions in 1999 (77% Mariaville, 85% Duanespurg). The Mariaville lake model includes SSTs (PC1) and pre-season chlorophyll-*a* as predictors, whereas the Duansespurg model includes only SSTs (PC2). The performance of the Duanespurg model compared to the Mariaville model again suggests that while global processes are important in setting conditions for peak season algae biomass, and both explain significant variability in the *magnitude* timeseries of both lakes, within-lake processes that may determine interannual variability of peak season algae abundance are not entirely captured by the hydroclimatic predictors investigated here, or by pre-season algae abundance. The variability in forecast skill and predictive power of hydroclimatic variables among study lakes highlights the importance of catchment- and lake-specific processes and characteristics in determining the effects of external climate forcing on peak-season algae abundance. Catchment soil types, land use, and location, as well as lake area, depth, and management may all influence the susceptibility of lake systems to climate variables (Moss, 2012), and may alter predictive skill.

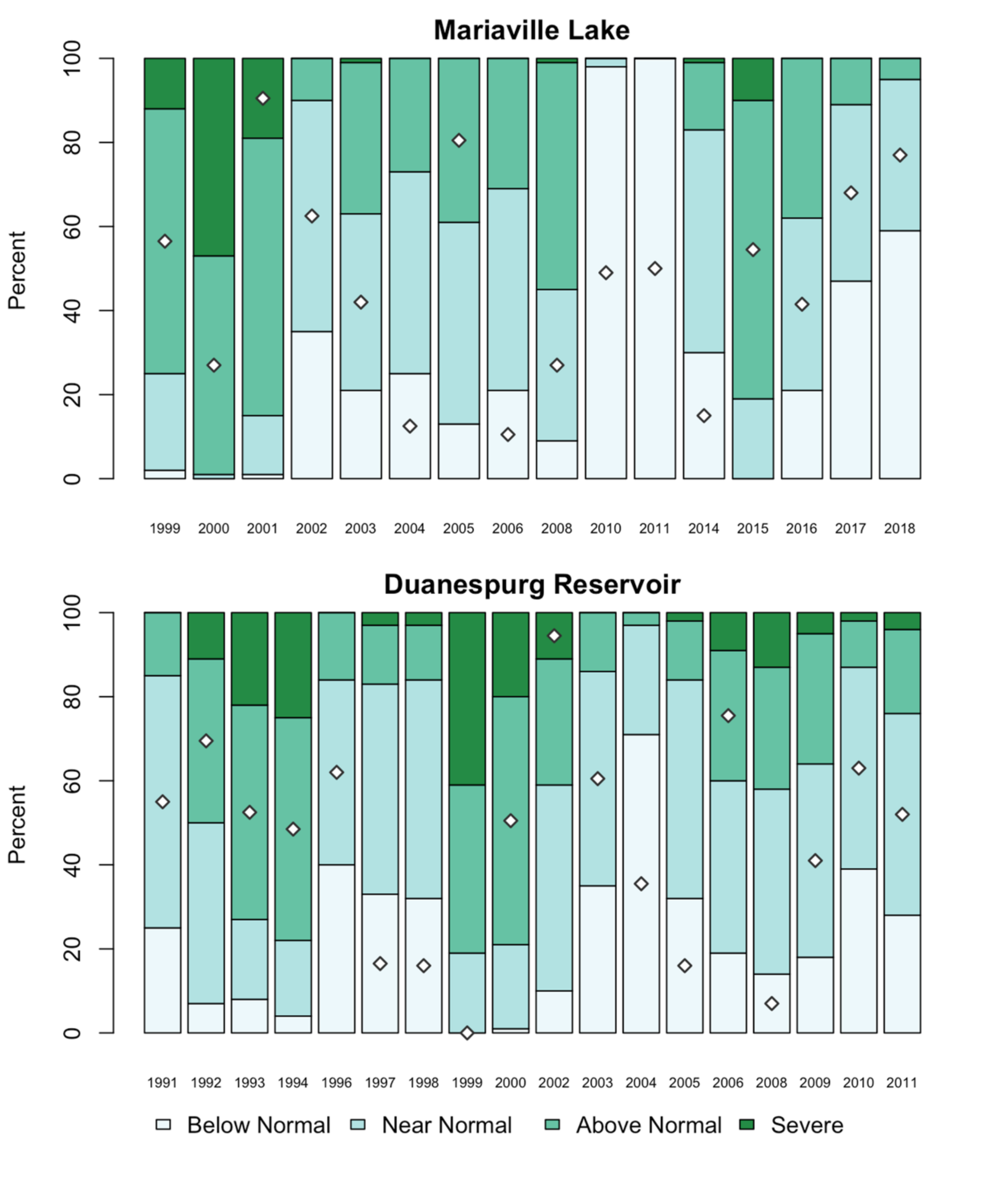

Figure 3-10. Probabilities of *magnitude* prediction categories for Mariaville Lake and Duanespurg reservoir, New York. Diamonds indicate the observed category.

While forecasting models developed on the selected pre-season predictors cannot entirely capture the nuances of peak season algae biomass, it is notable that relatively simplistic statistical models based on global sea surface temperatures and pre-season chlorophyll-*a* show significant skill in many of the selected lakes. For these lakes, season-ahead prediction of algae metrics may provide actionable information to lake managers and public health officials based on easily accessible gridded datasets and basic water quality monitoring.

3.5 Conclusions

In this chapter, season ahead predictions for July-October algae *magnitude* and *duration* are developed for 135 lakes identified across the Northeast and Midwest U.S. to inform lake management decisions prior to peak algae biomass. Prediction models are conditioned on local and global scale pre-season (March-June), readily available, gridded hydroclimatic variables and pre-season chlorophyll-*a*. Global SST and pre-season chlorophyll-*a* are the most common sources of predictive power across lake models. SST grids selected for prediction model development are concentrated in the northern Atlantic and equatorial Pacific, with characteristics of both ENSO and NAO.

Forecasting models outperform climatology in 87% (70%) of *magnitude* models and in 96% (94%) of *duration* models based on HSS (RPSS). Additionally, skillful prediction of elevated algae metrics, based on *magnitude* and *duration*, is evident in more than half of the included lakes. As cultural eutrophication fuels an expansion of harmful algae in lakes across the U.S., prediction tools to inform water quality management, particularly those conditioned on easily accessible data, may incentive preparedness actions and lake management decisions toward protecting public health and informing recreational activities.

Chapter 4. A Machine Learning and Remote Sensing-based Model for Algae Pigment and Dissolved Oxygen Retrieval on a Small Inland Lake

Adapted from: Beal, M. R., Özdoğan, M., & Block, P. J. (2024). A machine learning and remote sensing-based model for algae pigment and dissolved oxygen retrieval on a small inland lake. Water Resources Research, 60(3), e2023WR035744. DOI: 10.1029/2023WR035744

4.1 Introduction

In recent decades, many waterbodies have experienced an increase in eutrophication and the prevalence of harmful algae, resulting from widespread disturbance of phosphorus and nitrogen cycles, tied to large scale land cover and climate change (O'Neil et al., 2012; Paerl & Huisman, 2008; Paerl & Paul, 2012). Excessive algae growth can lead to negative consequences for ecosystem function (Huisman et al., 2018), economic opportunity (Dodds et al., 2009), and human and animal health (Carmichael & Boyer, 2016), due to the capability for toxin production in certain species. Cyanobacteria (blue-green algae) are a species of photosynthetic microorganisms of particular concern in freshwater systems (Paerl et al., 2001), that are capable of producing a range of toxins (Carmichael, 1994, 2001). Rapid growth of cyanobacteria may result in cyanobacterial harmful algal blooms (cHABs). Given the threats posed by cHABs, water managers urgently need novel techniques to monitor and manage cyanobacteria in freshwater systems.

Due to the cost-effectiveness and regular availability of satellite image data, remote sensing has become a powerful tool for water quality monitoring (Yan et al., 2018). In remote sensing, chlorophyll-a and phycocyanin are often used as surrogates for cyanobacteria abundance, chlorophyll-a representing all phytoplankton and phycocyanin being characteristic of cyanobacteria (Dekker, 1993). Both chlorophyll-a and phycocyanin may be useful from a management perspective. Stumpf et al (2016), found both chlorophyll-a and phycocyanin to be useful in statistical modeling for retrieval of cyanobacterial toxins. Additionally, the ability to discriminate between cyanobacteria and other algae species during a bloom event may allow water managers to make informed decisions about closing beaches or communicating water quality information to the public. Satellite-based estimates of algae indicators may also supplement more costly in-situ sampling efforts.

In addition to algal pigments, there is potential for retrieval of non-optical water quality variables via satellite imagery. Dissolved oxygen (DO) is an important indicator of water quality and can be strongly tied to phytoplankton dynamics in lakes (Qin et al., 2013). Phytoplankton increase DO concentrations through photosynthesis and consume DO during cellular respiration and decomposition. Sufficient DO concentrations are required to support many aquatic organisms, and abnormally low concentrations can result in distress or death of fish (Magee et al., 2019; Swingle, 1968). While direct relationships between water reflectance and DO have shown low potential for DO monitoring (Gholizadeh et al., 2016), relationships between DO and other remotely-sensed water quality parameters have shown some promise (Kim et al., 2020). If in situ algae metrics are statistically related to DO, and can be adequately acquired from remotely sensed imagery, indirect estimates of DO concentrations may be possible. Such DO pseudo-observations may provide additional information in identifying algae growth (high DO) and deoxygenated areas that may suggest a risk to aquatic life.

Phycocyanin and chlorophyll-a have similar spectral signatures, which makes differentiation difficult from a remote sensing perspective. Phycocyanin absorbs strongly at 620 nm, while chlorophyll-a absorbs between 665-681 nm (Stumpf, Davis, et al., 2016). Satellite imagery from several missions have been used to identify algae blooms in recent decades. Landsat, MODIS (Moderate Resolution Imaging Spectroradiometer), and MERIS (Medium Resolution Imaging Spectrometer) were found by Shi et al (2019) to be the most widely used products for monitoring cHABs. The Landsat series' long temporal record (~30 years) and fine spatial resolution (30m) provides a large and detailed imagery dataset frequently used for algae bloom monitoring (Boucher et al., 2018; Cao et al., 2020; Han & Jordan, 2005; Vincent et al., 2004). Similarly, the MODIS instrument, launched in 1999, has been used widely for water quality monitoring and identification of HABs (Becker et al., 2009; Binding et al., 2012). MODIS has a spatial resolution ranging between 250m and 1km, but a short revisit time (~one day) providing frequent observations useful for bloom monitoring. The MERIS instrument had a 10-year lifespan from 2002 – 2012. MERIS data has a 300m resolution and provides the best spectral resolution for monitoring inland water when compared to Landsat and MODIS (Shi et al., 2019). Unlike Landsat and MODIS, MERIS has bands located in spectral regions specific to both chlorophyll-a and phycocyanin making it well suited for detailed HAB characterization (Qi et al., 2014; Simis et al., 2005, 2007).

While several remotely sensed products exist for monitoring harmful algae, this work focuses on estimation of harmful algae metrics using Sentinel-2 and Sentinel-3 imagery. The Ocean and Land Color Instrument (OLCI) onboard the Sentinel-3 satellite is well suited for phycocyanin and chlorophyll-a retrieval, having sensors for bands located at 620 nm and between 665-681 nm, and a near daily revisit time. Additionally, the OLCI was developed as a direct successor to the MERIS instrument, potentially allowing for methods developed on OLCI data to be applied to historical MERIS imagery. The Multi Spectral Instrument (MSI) onboard the Sentinel-2 satellite does not have the spectral resolution or revisit time (~8 days for Sentinel-2) of Sentinel-3 but has a fine spatial resolution for land and water bands ranging from 10-20 m compared to the 300 m spatial resolution of OLCI. This high spatial resolution may be beneficial for developing cHAB monitoring tools on small inland waterbodies. Sentinel-2's MSI data has also been shown to be compatible with Landsat 8's OLI, potentially allowing for application of Sentinel-2 based water quality retrieval methods to a larger imagery dataset (Pahlevan et al., 2019).

In recent years, machine learning techniques have been successful in retrieving water quality parameters from satellite imagery (Silveira Kupssinskü et al., 2020; Su et al., 2021). Given the potential benefits of Sentinel-3's OLCI and Sentinel-2's MSI for water quality monitoring, the aim of this work is twofold: (1) to evaluate the ability of each instrument to retrieve chlorophyll-a, phycocyanin, and the phycocyanin: chlorophyll-a ratio (Pc:Chla below) from a small inland lake using machine learning methods, and (2) to develop a novel machine-learning based approach for indirect satellite-based estimations of dissolved oxygen conditioned on algae pigment concentrations.

4.2 Materials and Methods

4.2.1 Study Site and Water Quality Sampling

Despite its chronic water quality problems, little work has been done to test the efficacy of remotely sensed water quality monitoring methods on Lake Mendota. The workflow for development of remote sensing models has four main steps (1) in-situ sampling of water quality parameters during satellite overpasses, (2) downloading and processing relevant Sentinel-2 and Sentinel-3 imagery, (3) developing and validating machine learning models for estimation of algal pigments, and (4) development and validation of a machine learning model for indirect estimations of dissolved oxygen (Figure 4-1).

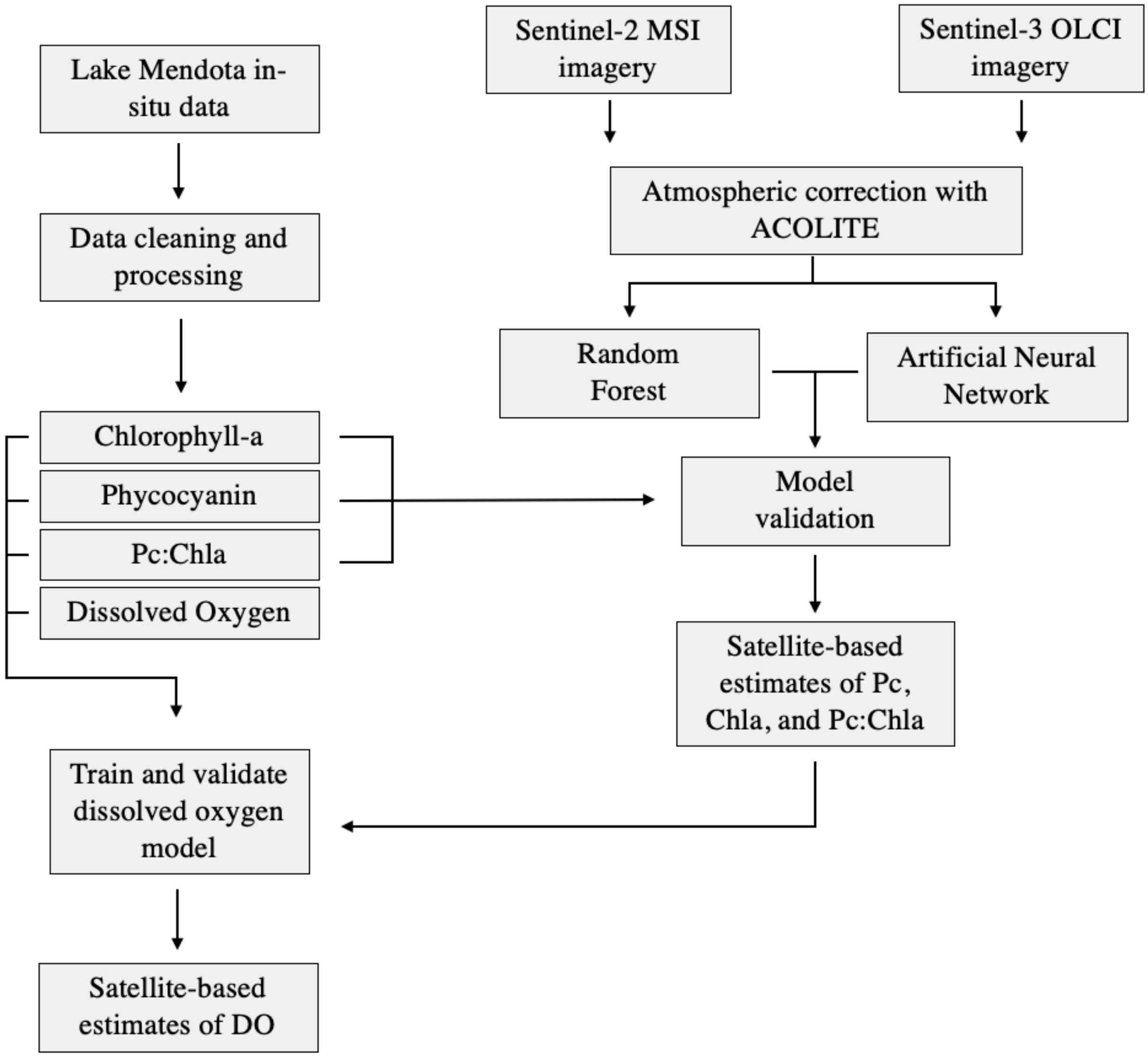
Lake Mendota in-situ data
Data cleaning and processing
Chlorophyll-a
Phycocyanin
Pc:Chla
Dissolved Oxygen
Train and validate dissolved oxygen model
Satellite-based estimates of DO
Sentinel-2 MSI imagery
Sentinel-3 OLCI imagery
Atmospheric correction with ACOLITE
Random Forest
Artificial Neural Network
Model validation
Satellite-based estimates of Pc, Chla, and Pc:Chla

Figure 4-1. Flowchart describing remote sensing and machine learning workflow.

To accurately represent spatial heterogeneity in Lake Mendota's water quality, phycocyanin and chlorophyll-a measurements were taken at 35 locations across Mendota using a YSI EXO II Sonde (Figure 4-2). To capture spatial variability in water quality and surface reflectance, points are sampled at random locations within 35 grid boxes (Figure 4-2). Point locations are randomized within each grid box before each sampling effort. This approach captures an array of sampling locations but can still characterize regions of the lake (grids) over time. At each point the sonde was used to collect data in surface waters for two minutes. The sonde was allowed to calibrate for the first minute, and the final data used in analysis is averaged over the second minute at each point. Phycocyanin and chlorophyll-a measurements were recorded in Relative Fluorescence Units and converted to micrograms per liter (µg/L) using a linear transformation provided by the sonde. The Pc:Chla ratio is calculated at each sampling point by dividing the two quantities. DO values were also collected with the sonde and reported in mg/L. Sampling efforts were conducted approximately once per week during the summer (typically June-August) from 2019 – 2022. Starting in 2021, water quality samples were also taken at the site of the UW-Madison NTL-LTER water quality monitoring buoy. Sampling efforts coincided with satellite overpasses from Sentinel-2 and Sentinel-3. Over four summers, 34 sampling trips were conducted generating 671 data points in total (Figure 4-3).

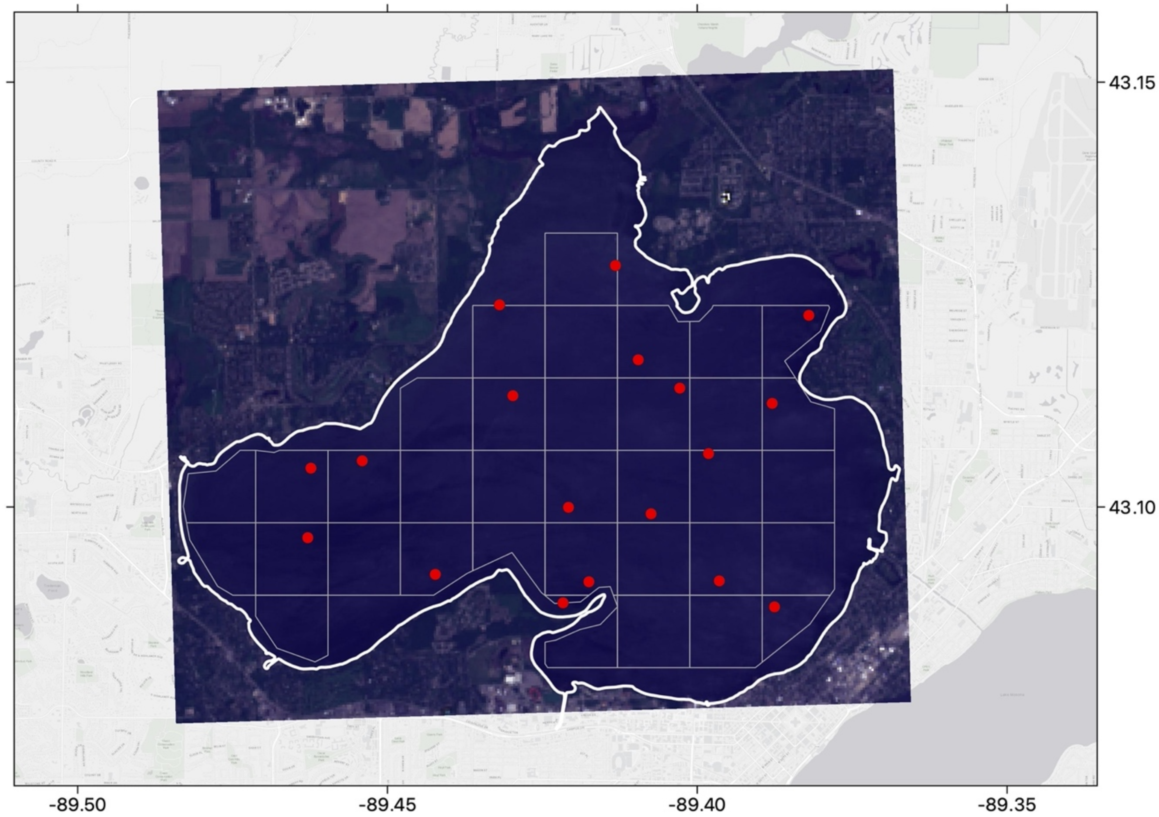

Figure 4-2. Lake Mendota sampling campaign on 2021-09-25 coinciding with a Sentinel-2 overpass. The Sentinel-2 image is pictured in real color at 60m^2 resolution. Red dots represent sample locations.

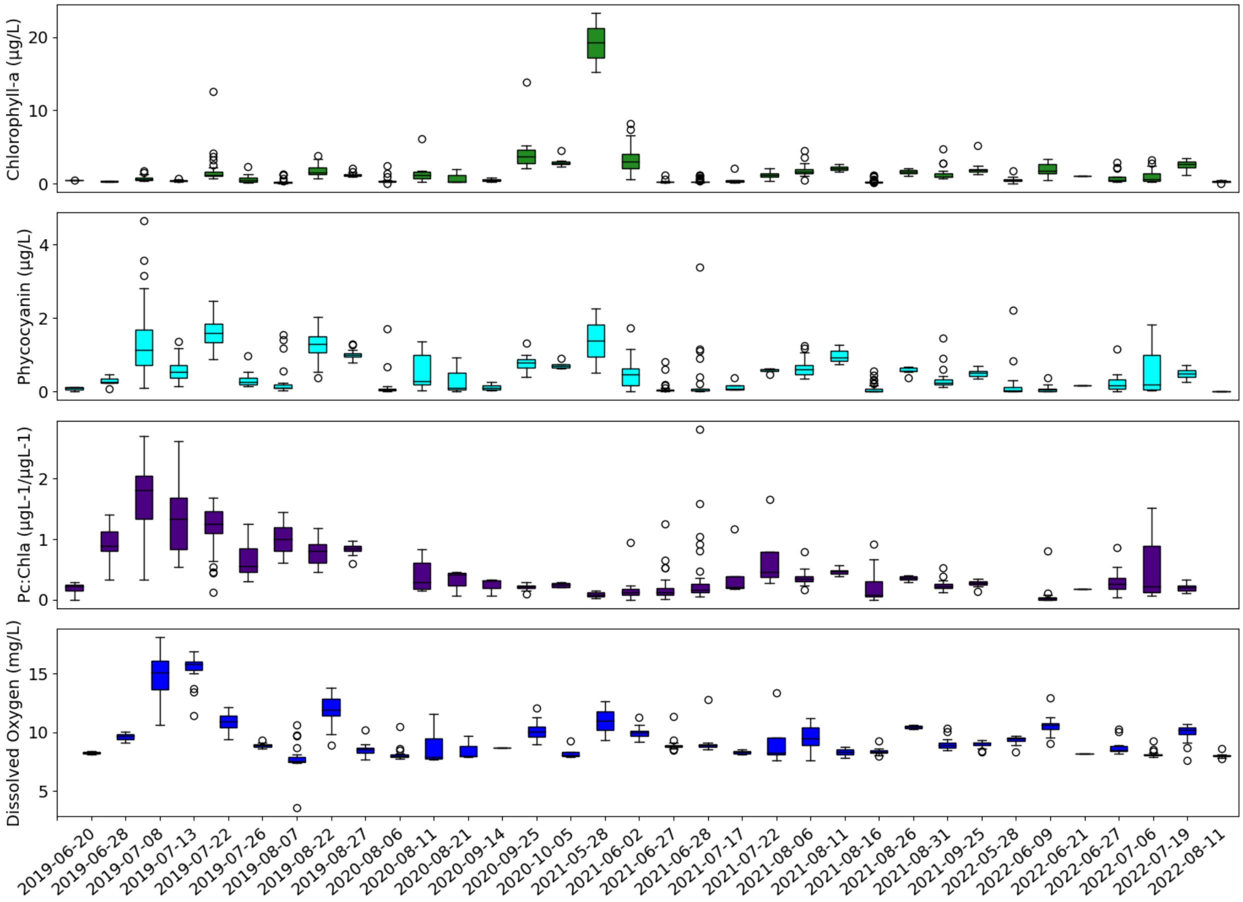

Figure 4-3. In situ chlorophyll-a, phycocyanin, Pc:Chla, and dissolved oxygen observations from sampling trips over the summers of 2019-2022.

4.2.2 Satellite Data

Sentinel-2 MSI and Sentinel-3 OLCI data are both made available by Copernicus (Copernicus, 2022). Sentinel-2 and Sentinel-3 top-of-atmosphere (TOA) full resolution radiance images are obtained directly from the Copernicus Open Access Hub and processed with ACOLITE, an image processor developed for atmospheric correction and processing of satellite images for coastal and inland water applications (Vanhellemont & Ruddick, 2018, 2021), to provide surface reflectance values (Table 4-1). Sentinel-2 images are resampled to 60m.

Table 4-1. Atmospheric correction results.

Satellite	Band	Top of Atmosphere		ACOLITE	
		Mean	Standard Deviation	Mean	Standard Deviation
Sentinel 2	B1	0.43	0.28	0.018	0.011
	B2	0.41	0.29	0.024	0.011
	B3	0.38	0.27	0.044	0.021
	B4	0.38	0.31	0.020	0.014
	B5	0.39	0.31	0.024	0.016
	B6	0.39	0.31	0.017	0.015
	B7	0.40	0.32	0.018	0.015
	B8	0.40	0.33	0.015	0.014
	B8A	0.21	0.31	0.014	0.016
	B11	0.21	0.08	0.005	0.015
	B12	0.18	0.20	0.007	0.013
Sentinel 3	Oa01_radiance	-1.0E+07	1.1E+07	0.05	0.04
	Oa02_radiance	1.95	1.20	0.01	0.00
	Oa03_radiance	1.89	1.24	0.04	0.04
	Oa04_radiance	1.80	1.30	0.04	0.04
	Oa05_radiance	1.51	1.17	0.04	0.03
	Oa06_radiance	1.69	1.35	0.05	0.03
	Oa07_radiance	1.05	0.89	0.07	0.03
	Oa08_radiance	0.96	0.86	0.04	0.03
	Oa09_radiance	1.01	0.93	0.03	0.03
	Oa10_radiance	-5.8E+06	5.9E+06	0.03	0.03
	Oa11_radiance	0.69	0.63	0.01	0.01
	Oa12_radiance	0.69	0.64	0.03	0.03
	Oa13_radiance	0.49	0.54	0.04	0.03
	Oa14_radiance	0.37	0.40	0.04	0.04
	Oa15_radiance	0.35	0.35	0.04	0.04
	Oa16_radiance	0.44	0.41	0.04	0.04
	Oa17_radiance	0.39	0.36	0.03	0.04
	Oa18_radiance	0.38	0.36	0.03	0.04
	Oa19_radiance	0.30	0.29	0.02	0.02
	Oa21_radiance	0.13	0.14	0.03	0.04

4.2.3 Machine Learning Approach

A machine learning regression approach is adopted to model chlorophyll-a, phycocyanin, phycocyanin/chlorophyll-a. Models are developed separately for Sentinel-2 and Sentinel-3 images. Multiple machine learning architectures have shown promise in the retrieval of algae metrics and categorization of algal blooms using remotely sensed data (Hill et al., 2020). In this study, a random forest (RF) regression model and an artificial neural network (ANN) are trained on Sentinel-2 MSI and Sentinel-3 OLCI data. The random forest model is implemented using the sklearn package in python; the artificial neural network model leverages the Keras package.

Random forests are created with a collection of tree-structured classifiers, generated by a random selection of training data, that collectively "vote" on an outcome (Breiman, 2001). RFs are often implemented as a regression in which the outcomes of individual trees are averaged to generate a continuous value result. This approach has become a commonly used regression and classification tool in remote sensing due its explanatory power, ability to select and rank the importance of input variables, and minimal tuning requirements (Belgiu & Drăguţ, 2016).

Artificial Neural Networks are constructed using a collection of neurons and edges with associated weights. The weights are adjusted as the ANN processes training data, learning the relationship between inputs and outputs. These models have been used effectively to model the complex relationships between image reflectance and water quality parameters, but often require computationally expensive parameter tuning to achieve skillful results (Chebud et al., 2012). ANNs developed here use the rectified linear unit activation function and adam, a stochastic gradient descent optimization method. ANNs are tuned to determine a satisfactory number of epochs, batch size, and weight initializer using the scikit-learn package (Pedregosa et al., 2011). Uniform and normal initializers are tested.

To increase the effectiveness of machine learning models and identify the most important spectral features associated with phycocyanin and chlorophyll-a, several band combinations are included in the suite of model inputs (explanatory variables), in addition to each of the OLCI and MSI bands (Table 4-2). Many of these algorithms are taken from Beck et al. (2017), who generalize a number of satellite-specific chlorophyll-a and phycocyanin detection algorithms for use across a range of satellites, including Sentinel-2 and Sentinel-3. Algorithms modified by Beck et al. (2017) are credited to the original authors. To assess the importance of different bands and band algorithms, the Boruta variable selection approach is applied during the building of RF models (Kursa et al., 2010). Boruta variable selection randomly shuffles each feature and tests original input variable against randomized ones in an iterative process. Features that are less relevant than random variables are removed from the set of inputs. Features are selected separately in each fold; the final model uses the features that are chosen in every fold-specific selection process. For consistency, the input variables selected for the RF are also used in the development of ANN models.

It should be noted that Sentinel-2 cannot sense the 620 nm phycocyanin absorption feature commonly used in phycocyanin retrieval, however, several studies have found success in using Sentinel-2 based chlorophyll-a proxy algorithms for indirect phycocyanin estimation (Beck et al., 2017; Pérez-González et al., 2021; Sòria-Perpinyà et al., 2020). Therefore, a combination of proxy algorithms, machine learning techniques, and feature selection methods is used to develop and assess models for phycocyanin and chlorophyll-a retrieval based on Sentinel-2 imagery.

In order to assess each satellite's ability to retrieve DO concentrations, a non-optical water quality variable, RF and ANN models are constructed to assess any relationships between DO and each of the in situ algae indicators: chlorophyll-a, phycocyanin, and Pc:Chla. The models' ability to indirectly estimate DO are then assessed using remotely sensed pseudo-observations of chlorophyll-a, phycocyanin, and Pc:Chla. Model errors are evaluated spatially for both satellites.

Table 4-2 Band combinations included as input variables for phycocyanin, chlorophyll-*a*, and Pc:Chla by satellite. Rrs indicates Remote sensing reflectance.

Satellite	Reference	Algorithm
Sentinel-2	Amin et al. (2009)	Rrs(705)-Rrs(665)
MSI	Amin et al. (2009)	Rrs(705)/Rrs(665)
	Beck et al. (2017)	Rrs(740)-Rrs(665)
	Beck et al. (2017)	Rrs(740)/Rrs(665)
	Beck et al. (2017)	Rrs(740)-Rrs(705)
	Beck et al. (2017)	Rrs(740)/Rrs(705)
	Beck et al. (2017)	Rrs(740)-Rrs(665)/Rrs(740)+Rrs(665)
	Beck et al. (2017)	Rrs(740)-Rrs(705)/Rrs(740)+Rrs(705)
	Gitelson et al. (2003)	(1/Rrs(705)) – (1/Rrs(665)) – Rrs(740)
	Hu (2009)	Floating Algae Index
	Beck et al. (2016)	Rrs(560)-[Rrs(665)+(Rrs(490)-Rrs(665))]
	Mishra and Mishra (2012)	[Rrs(709)-Rrs(665)]/[Rrs(708)+Rrs(665)]
	Gons et al. (2002)†	$R_m(a_w(709) + bb) - aw(665) - b_b{}^p\}/a^*(665)$
	Moses et al. (2012)	[113.36 * {[Rrs(665)^-1-Rrs(709)^-1]*Rrs(753)}+16.45]^1.124
Sentinel-3	Alawadi et al. (2010)	Rrs(865)-Rrs(665)/Rrs(443)+Rrs(510)
OLCI	Amin et al. (2009)	Rrs(681)-Rrs(620)
	Amin et al. (2009)	Rrs(681)-Rrs(665)/Rrs(681)+Rrs(665)
	Beck et al. (2016)	Rrs(560)-[Rrs(665)+Rrs(443)-Rrs(665)]
	Beck et al. (2017)	Rrs(620)-[Rrs(709+(Rrs(560)-Rrs(709))]
	Beck et al. (2017)	Rrs(709)-Rrs(620)
	Gower et al. (2004)	[Rrs(560)-Rrs(681)]-[Rrs(754)-Rrs(681)]
	Gitelson et al. (2003)	(1/Rrs(620)) – (1/Rrs(560)) – Rrs(709)
	Kneubuhler et al. (2007)	[Rrs(443)-Rrs(665)]/Rrs(510)
	Mishra et al. (2009)	Rrs(709)/Rrs(681)
	Beck et al. (2017)	Rrs(709)-Rrs(620)/Rrs(709)+Rrs(620)
	Mishra and Mishra (2014)	(1/Rrs(709)) – (1/Rrs(665)) * Rrs(709)
	Simis et al. (2005)	Rrs(709)/Rrs(620)
	Schalles and Yacobi (2000)	Rrs(665)/Rrs(620)
	Stumpf et al. (2016)	[Rrs(665)-Rrs(620)]+[Rrs(620)-Rrs(681)*0.74]
	Wynne et al. (2008)	-1*Rrs(681)-Rrs(620)-[Rrs(709)-Rrs(620)]
	Mishra and Mishra (2012)	[Rrs(709)-Rrs(665)]/[Rrs(708)+Rrs(665)]
	Gons et al. (2002)†	$R_m(a_w(709) + bb) - a_w(665) - b_b{}^p\}/a^*(665)$
	Moses et al. (2012)	[113.36 * {[Rrs(665)^-1-Rrs(709)^-1]*Rrs(753)}+16.45]^1.124

† The algorithm from Gons et al. (2002) is semi-analytical, parameters are derived by Acolite during processing.

4.3 Results

4.3.1 Sentinel-2 MSI

Sentinel-2 overpasses coincided with 10 sampling dates between 2019 and 2022. After processing images and masking clouds, this leaves 206 points of in situ data for comparison with Sentinel-2 MSI reflectance data (Figure 4-4). On several sampling days there are two Sentinel-2 images over Lake Mendota. These additional images were removed during model construction to minimize artificial inflation of model skill, however inclusion resulted in insignificant differences.

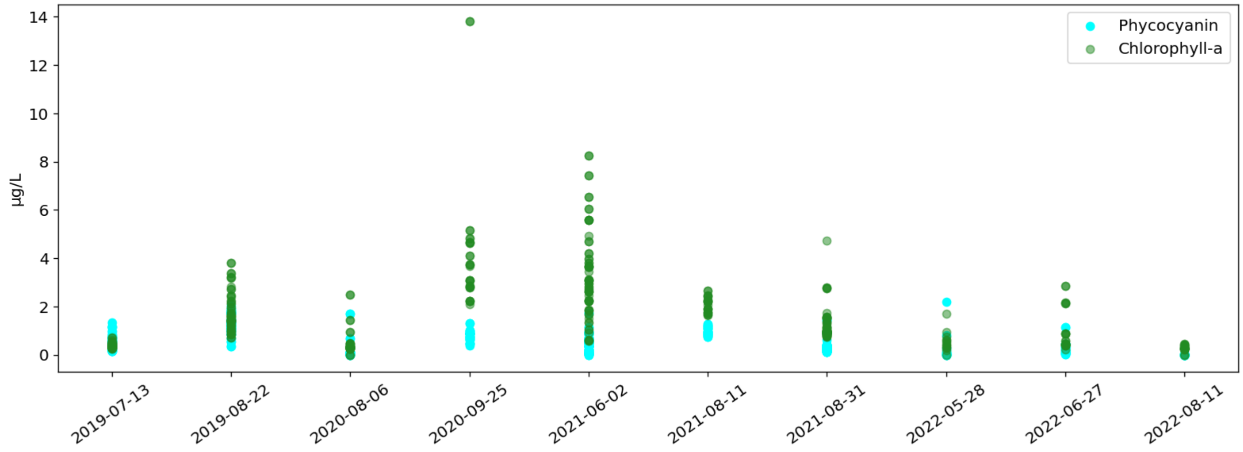

Figure 4-4. In situ chlorophyll-a and phycocyanin data on Sentinel-2 overpass days.

RF and ANN regressions are built using a 5-fold cross-validation approach; in iterative fashion, the model is constructed on 80% of the available data and evaluated on the remaining 20% of data and repeated five times. Variable selection is applied in each fold when constructing the RF model. Variables retained for each Sentinel-2 model are listed in Table 4-3. The chlorophyll-a model retains two bands and three algorithms. Rrs(443) targets chlorophyll-a absorption maximum. Rrs(560) is one of the reflectance peaks for chlorophyll-a. The two subtraction algorithms make use of the chlorophyll-a absorption peak (Rrr(665)), and Rrs(705) and Rrs(740), two near infrared bands often used for identifying vegetation. The final algorithm is developed by Beck et al. (2016)

to target chlorophyll-a concentrations by evaluating differences between spectral signatures of clear water (Rrs(490)), chlorophyll-a absorption, and chlorophyll-a reflection (Rrs(560)). The phycocyanin model includes similar input variables to the chlorophyll-a model, with the exception of Rrs(740)-Rrs(705), and adds an algorithm created by Gitelson et al. (2003) that was originally developed to estimate leaf chlorophyll-a concentrations, and was later adapted to water quality applications by Beck et al. (2016). The Pc:chla model retains variables found in the chlorophyll-a and phycocyanin models, but adds an algorithm focusing on the differences between Rrs(740) and Rrs(705).

Table 4-3 Bands and algorithms retained for Sentinel-2 models.

Model	**Variables selected**
Chlorophyll-a	Rrs(443)
	Rrs(560)
	Rrs(740)-Rrs(705)
	Rrs(740)-Rrs(665)
	Rrs(560)-[Rrs(665)+(Rrs(490)-Rrs(665))]
Phycocyanin	Rrs(443)
	Rrs(560)
	Rrs(705)
	Rrs(740)-Rrs(665)
	Rrs(740)-Rrs(705)
	Rrs(705)-Rrs(665)
	Rrs(560)-[Rrs(665)+(Rrs(490)-Rrs(665))]
	(1/Rrs(705)) – (1/Rrs(665)) – Rrs(740)
Pc:Chla	Rrs(560)
	Rrs(705)
	Rrs(740)-Rrs(665)
	Rrs(740)/Rrs(665)
	Rrs(560)-[Rrs(665)+(Rrs(490)-Rrs(665))]
	Rrs(740)-Rrs(705)/Rrs(740)+Rrs(705)
	(1/Rrs(705)) – (1/Rrs(665)) – Rrs(740)

All models are notably skillful (Figure 4-5, Table 4-4). Phycocyanin and Pc:Chla models perform better than models for chlorophyll-a. RF models outperform ANN models across all variables by fold averaged coefficient of determination (R^2). Fold averaged mean absolute error (MAE) is slightly lower for the chlorophyll-a ANN compared to the RF. While the chlorophyll-a models have similar performance scores, the ANN model appears to better capture the higher chlorophyll-a values, which may be desirable from a monitoring perspective. Phycocyanin models generally perform well across the range of observed values. Both chlorophyll-a and phycocyanin models appear to have some trouble capturing zero values but can separate low and high concentrations overall. The Pc:Chla model appears to perform best at low ratio values (0-1), which may suggest the model is better at identifying high chlorophyll-a concentrations. Strong performance in lower Pc:Chla values likely contributes to the overall high skill. The model is also accurate for several high Pc:Chla ratios, suggesting some ability to differentiate between blue-green and green blooms.

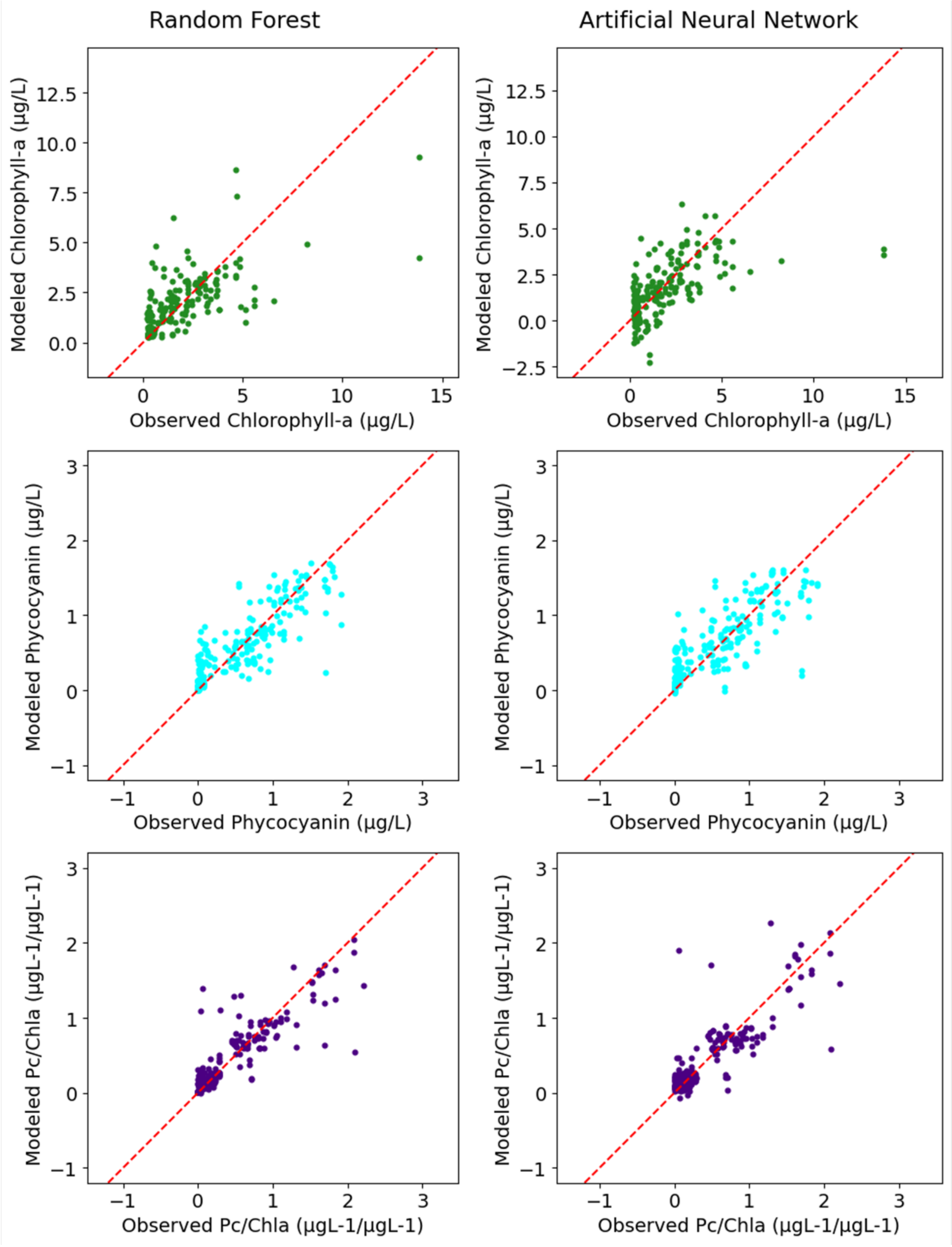

Figure 4-5. Observed vs. modeled phycocyanin, chlorophyll-*a*, and phycocyanin/chlorophyll-*a* for RF and ANN models with Sentinel-2 MSI data. The red line represents a 1:1 slope.

Table 4-4 Sentinel-2 fold averaged coefficient of determination (R^2) and mean absolute error for RF and ANN models.

Algae Metric	Random Forest		Artificial Neural Network	
	R^2	MAE (μg/L)	R^2	MAE (μg/L)
Chlorophyll-a	0.47	0.81	0.44	0.78
Phycocyanin	0.69	0.21	0.58	0.24
Phycocyanin/Chlorophyll-a	0.70	0.14	0.42	0.18

4.3.2 Sentinel-3 OLCI

Sentinel-3 overpasses coincided with 11 sampling dates. After processing images and masking clouds, 161 points of in situ data were available for comparison with OLCI reflectance data (Figure 4-6). RF and ANN models for Sentinel-3 data apply the same 5-fold cross validation approach and boruta variable selection process as the Sentinel-2 models. Variables selected for each Sentinel-3 model are listed in Table 4-5.

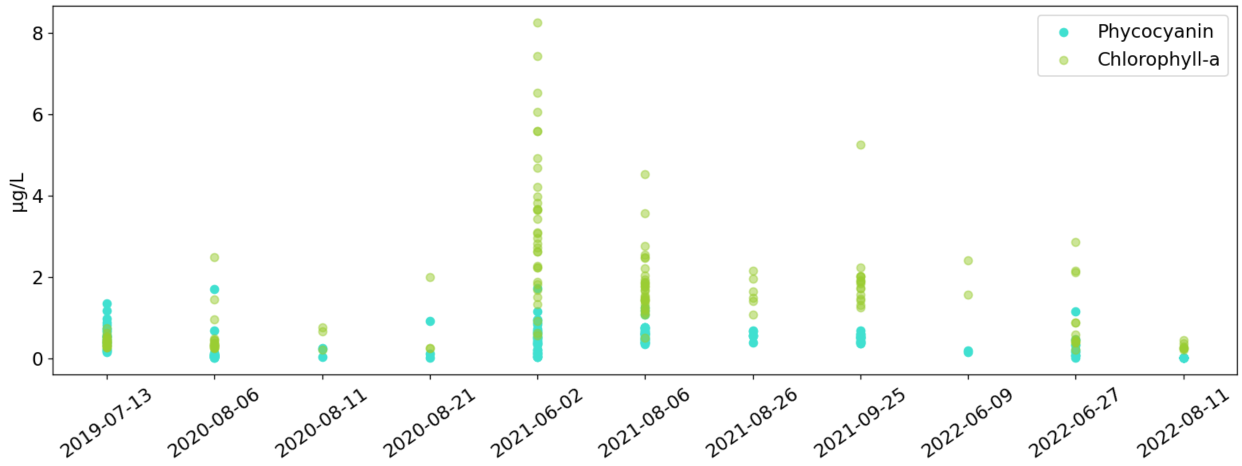

Figure 4-6. In situ chlorophyll-a and phycocyanin data on Sentinel-3 overpass days.

The Sentinel-3 chlorophyll-a model retains three variables in all 5 folds including two bands and one algorithm. Rrs(560) is the reflectance peak for chlorophyll-a. The algorithms retained were developed by Beck et al. (2016, 2017) to target chlorophyll-a and phycocyanin concentrations respectively. The chlorophyll-a algorithm estimates concentrations using the difference between the chlorophyll-a reflectance peak and the two absorption maxima: (Rrs(665) and Rrs(443)). The phycocyanin algorithm uses the difference between the phycocyanin absorption maximum (Rrs(620)) and chlorophyll-a reflectance. The Sentinel-3 phycocyanin model retains two bands and one algorithm. Rrs(412) is often used to measure turbidity, while Rrs(443) targets a chlorophyll-a absorption maximum. The algorithm retained is developed by Alawadi (2010) to target floating algae in ocean environments. The phycocyanin model does not retain the Rrs(620) band which targets spectral features of phycocyanin. The absence of a relationship between Rrs(620) and phycocyanin concentrations may speak to the relatively poor performance of the Sentinel-3 phycocyanin model, discussed below. The Sentinel-3 Pc:Chla model retains three algorithms. The Pc:Chla model leverages algorithms from both the chlorophyll-a and phycocyanin models, and an additional algorithm focusing on the difference between the phycocyanin absorption peak and the chlorophyll-a fluorescence baseline (Rrs(709)).

Table 4-5 Bands and algorithms retained for Sentinel-3 models.

Model	Variables selected
Chlorophyll-a	Rrs(560)
	Rrs(560)-[Rrs(665)+Rrs(443)-Rrs(665)]
	Rrs(620)-[Rrs(709)+Rrs(560)-Rrs(709)]
Phycocyanin	Rrs(412)
	Rrs(443)
	Rrs(865)-Rrs(665)/Rrs(443)+Rrs(508)
Pc:Chla	Rrs(865)-Rrs(665)/Rrs(443)+Rrs(508)
	Rrs(620)-(Rrs(709)+Rrs(560)-Rrs(709)
	Rrs(709)-Rrs(620)

Sentinel-3 models demonstrate mixed skill. Generally, Sentinel-3 models have lower fold averaged R^2 scores than Sentinel-2 models but return similar mean absolute error values (Table 4-6).

Chlorophyll-a R^2 values are superior to phycocyanin and Pc:Chla. RF models perform slightly better than the ANN models and appear to capture higher chlorophyll-a values well (Figure 4-7). The Sentinel-3 chlorophyll-a model is comparable to the Sentinel-2 model, albeit with a lower mean R^2 but also lower mean absolute errors. The phycocyanin model has the worst performance of the Sentinel-3 models. Both the RF and ANN appear to struggle with extremely low and high concentrations; however, the ANN does demonstrate some improvement over the RF for concentrations above 1 μg/L. The Pc:Chla model shows skill similar to the chlorophyll-a model. Like Sentinel-2, model results for low Pc:Chla ratios appear strong, with much greater scatter in high values.

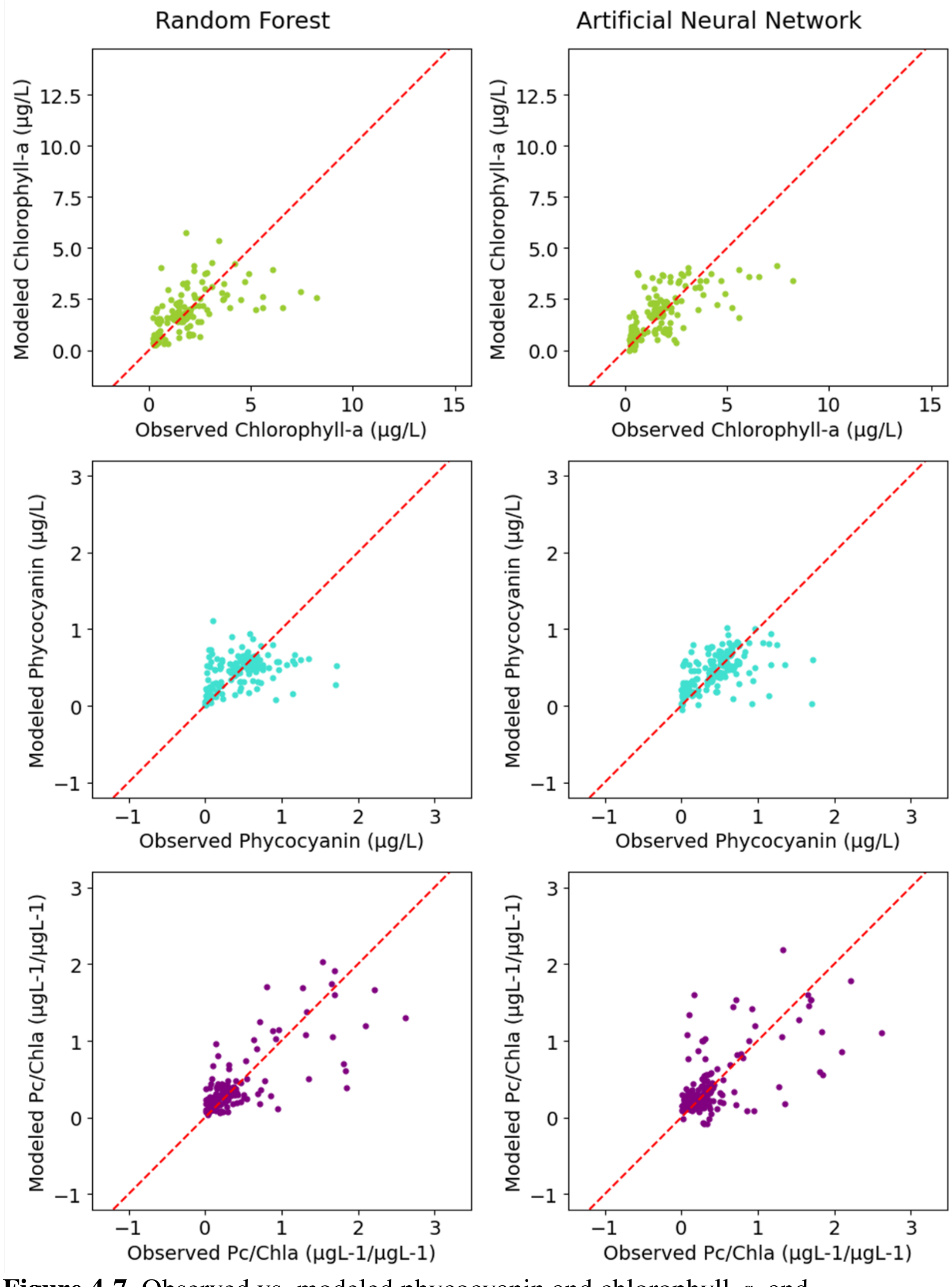

Figure 4-7. Observed vs. modeled phycocyanin and chlorophyll-*a*, and phycocyanin/chlorophyll-*a* for RF and ANN models with Sentinel-3 OLCI data. The red line represents a 1:1 slope.

Table 4-6 Sentinel-3 fold averaged coefficient of determination (R^2) and mean absolute error for RF and ANN models.

Algae Metric	Random Forest		Artificial Neural Network	
	R^2	MAE (μg/L)	R^2	MAE (μg/L)
Chlorophyll-a	0.42	0.72	0.41	0.74
Phycocyanin	0.09	0.22	0.22	0.20
Phycocyanin/Chlorophyll-a	0.41	0.18	0.31	0.23

4.3.3 Dissolved Oxygen

DO is modeled using the same 5-fold cross validation process and both RF and ANN model structures. Models are trained on all available in-situ data (671 points), including chlorophyll-a, phycocyanin, and Pc:Chla. Both in-situ models have notable skill. The ANN reports a fold averaged R^2 of 0.49 and a MAE of 0.97 mg/L. The RF model performs slightly better with a fold averaged R^2 of 0.53 and a MAE of 0.91 mg/L. Using the same model structure, but substituting in Sentinel-2 data (206 points), the RF (R^2: 0.68, MAE: 1.04 mg/L Figure 8) significantly outperforms the ANN (R^2: 0.36, MAE: 1.37 mg/L). To assess the transferability of the DO model, chlorophyll-a, phycocyanin, and Pc:Chla data at the Lake Mendota buoy are used to construct a similar DO model (Magnuson et al., 2023). Averaged hourly data between 6am - 6pm from 2019 to present was retained (N=10,675). 6am-6pm is chosen to best compare with spatial sampling data. Using 5-fold cross validation, the buoy ANN model results in a fold averaged R^2 of 0.53 and a MAE of 1.28 mg/L; the RF model results in a fold averaged R^2 of 0.50 and a MAE of 1.24 mg/L. None of the Mendota Buoy data was used in the construction of remote sensing models.

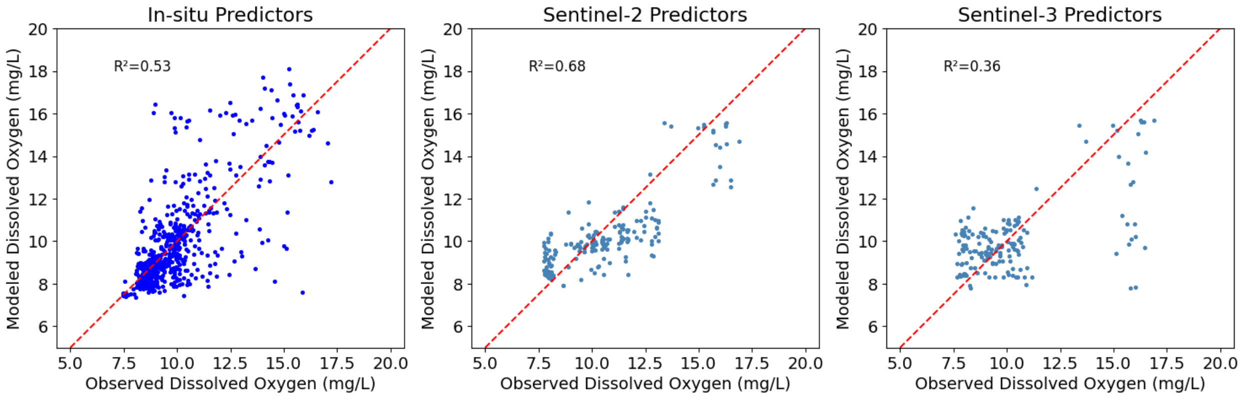

Figure 4-8. RF model results for dissolved oxygen based on in situ data (left) and applied to remotely sensed data (right).

By comparison, skill is notably lower with Sentinel-3 data for both the RF (R^2: 0.36, MAE: 1.37 mg/L; Figure 4-8) and ANN, which shows no skill (R^2: -1.4, MAE: 2.42 mg/L). The weaker performance of Sentinel-3 imagery in retrieving harmful algae indicators is evident in the performance of the DO model, which struggles to capture high DO concentrations.

4.3.4 Spatial Heterogeneity

To better understand the spatial performance of each satellite-based model, the mean absolute error is calculated using all available observed and modeled data pairs. While modeling of water quality parameters is performed at the pixel scale of each satellite, data from error estimates are aggregated to the scale of the 35 sampling grid boxes for comparison. For each water quality parameter, the best performing model type (ANN or RF) for each satellite is used. Variability is evident across the lake and between parameters (Figure 4-9).

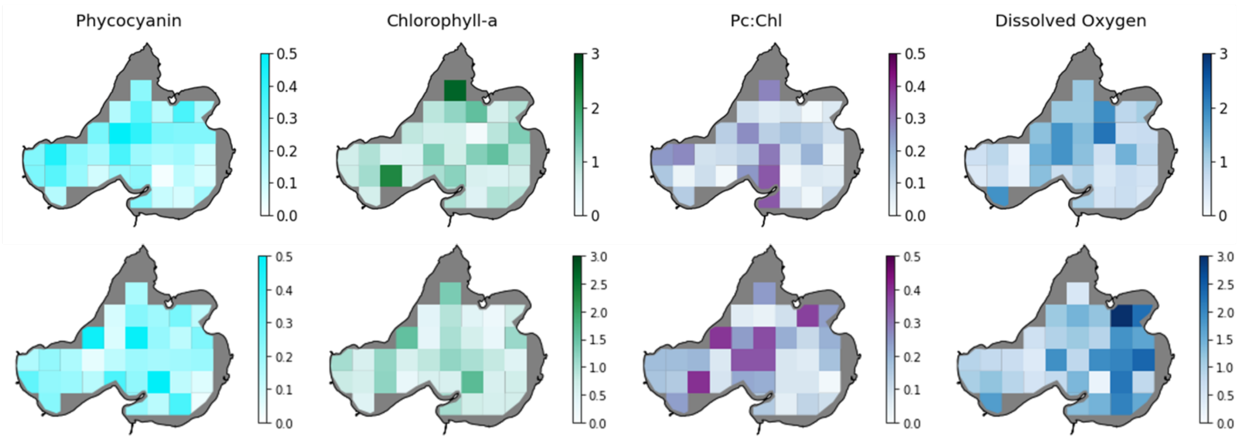

Figure 4-9. Mean absolute error between in-situ and Sentinel-2 (top) and Sentinel-3 (bottom) modeled water quality parameters.

Overall, the magnitude of errors in phycocyanin are similar between Sentinel-2 and Sentinel-3. Both satellites generally capture phycocyanin well across the lake, with ~90% of points falling below an MAE of 0.3 µg/L and a maximum MAE of 0.5 µg/L. For Sentinel-2, errors are highest along the northwestern shoreline of Mendota. For Sentinel-3, high MAE values are evident along the southern shoreline and the northwestern portion of the lake. Sentinel-3 chlorophyll-a spatial MAE values are generally lower than for Sentinel-2; the latter has MAE values at 2.69 µg/L near the Yahara inlet at the northern end of the lake. Comparatively, Sentinel-3 reports its highest MAE along the northwestern shore of the lake at 1.62 µg/L. 90% of MAE values for both satellites fall below ~1.3 µg/L. Errors in chlorophyll-a appear to have little spatial coherence, however, both satellites generate relatively large errors in the northernmost grid near the Yahara inlet. Differences between satellite performance for Pc:Chl retrieval are more apparent. Overall, Pc:Chl MAE is slightly larger for Sentinel-3 with a cluster of high error locations in the center of the lake, the western bay, and the northeastern shore, ranging from 0.35 – 0.40 µgL^{-1}/µgL^{-1}. The highest Pc:Chla errors from Sentinel-2 are located near the southern shoreline of the lake, ranging from 0.29 – 0.34 µgL^{-1}/µgL^{-1}. Similarly, the skill of DO model results varies between the two satellites. Sentinel-3

DO errors are clustered on the eastern shoreline of the lake, with the five highest MAE values ranging from 2.2 to 3.2 mg/L. Sentinel-2 DO errors appear highest in northern lake Mendota near the Yahara River inlet. Sentinel-2 MAE is significantly lower, with the five highest MAE values ranging from 1.5 – 2.1 mg/L.

4.4 Discussion

4.4.1 Model Performance

For most water quality parameters, RF models outperform ANN models using either Sentinel-3 OLCI or Sentinel-2 MSI data; the only exception is for phycocyanin using Sentinel-3. The consistent underperformance of the ANN models may be attributable to several causes. ANNs require significant amounts of training data and extensive tuning to be effective. In remote sensing applications, Maxwell et al. (2018) find that ANNs typically have high sensitivity to both training data size and parameter sets, particularly when compared to RFs. The dataset used to train Sentinel-2 and Sentinel-3 models may be too small for an ANN to adequately learn relationships between band reflectance and the water quality variables evaluated here. This may also make ANN models prone to overfitting. ANN performance may conceivably improve with further parameter tuning and access to more data, however, the simplicity and strong performance of RF models suggests that it is better suited to phycocyanin and chlorophyll-a retrieval, at least for Lake Mendota. Furthermore, the interpretability of RF models provides important insight into how different water quality variables are identified from each satellite. Given the relatively limited water quality data in this single system, we recommend the use of RF for estimation of water quality variables. The additional complexity offered by an ANN approach does not appear to be justified.

Robust spatial performance is an important attribute for remotely sensed water quality models. As discussed previously, notable variability in absolute terms and errors exists across the lake for all metrics. For phycocyanin and chlorophyll-a, both satellites struggle near the inlet of the Yahara River and along the northwestern shoreline; the Yahara inlet has the highest average chlorophyll-a value across the lake. Similarly, high phycocyanin values occur near the inlet and in the western bay of the lake. High productivity combined with inflow of sediments and dissolved organic matter likely make distinguishing chlorophyll-a difficult in this part of the lake, as separating spectral signals of chlorophyll-a and phycocyanin is notoriously difficult (Gholizadeh et al., 2016). Spatial distribution of Pc:Chla errors is notably different between the two satellites. Errors for Sentinel-3 are clustered in the center of the lake and include two locations with the highest average Pc:Chla values across the lake. This is not surprising given Sentinel-3's limited ability to retrieve high Pc:Chla values. Errors in model results for Pc:Chla are notably lower for Sentinel-2 than Sentinel-3, and higher errors appear clustered around the southern portion of the lake. This might suggest interference from shoreline or shallow water reflectance. DO errors also vary between satellites. Sentinel-2 errors are highest near the Yahara River inlet while Sentinel-3 errors are highest in the northeastern portion of the lake. Because the DO model is conditioned on algae indicators, one possible explanation for the high error rate is an abundance of submersed aquatic vegetation in both regions, which may influence DO concentrations in the absence of algae.

4.4.2 Satellite Suitability

Sentinel-2 and Sentinel-3 offer different advantages for water quality monitoring. Sentinel-3's short return period allows for continuous, near real-time monitoring of lake conditions, while the spatial resolution offered by Sentinel-2 is a powerful tool for monitoring the spatial distribution of harmful algae on small lakes. In this analysis, Sentinel-2 is found to be better equipped for algae indicator retrieval in Lake Mendota compared to Sentinel-3. Despite the higher overpass frequency of Sentinel-3, and the presence of the 620nm band specific to phycocyanin detection, models conditioned on Sentinel-2 imagery outperformed Sentinel-3 models for all variables. There are several potential reasons for the difference in capability: the fine spatial resolution of Sentinel-2 may offer a better estimate of reflectance values where in situ data were collected, particularly if algae conditions are variable across the lake. Additionally, as discussed previously, machine learning models are sensitive to the size of training datasets. The number of points available for training models based on Sentinel-2 imagery (206 points) exceeds the number of points available for Sentinel-3 (161 points).

Variable selection implemented during RF model construction provides insights into the algorithms and bands selected for water quality retrieval. Chlorophyll-a models showed several similarities in variable selection between the two satellites. Both models include bands located at 560nm and 665nm. Additionally, bands located at 705nm and 709nm are retained by the Sentinel-2 and Sentinel-3 models, respectively. All these bands and associated algorithms have been used consistently in chlorophyll-a retrieval among various satellite missions. The development of several popular algorithms focus on this spectral region, including the Maximum Chlorophyll Index (Gower et al., 2005), the Normalized Difference Chlorophyll Index (Mishra & Mishra, 2012), and many two band algorithms (Gilerson et al., 2010), among others. Furthermore, chlorophyll-a retrieval algorithms based in this red-NIR region have been shown to perform well in productive waters as they are less sensitive to absorption by colored dissolved organic matter (CDOM) when compared with algorithms focused in blue-green wavelengths (Gilerson et al., 2010). Given Lake Mendota’s hypereutrophic status, the retention of bands established as important in chlorophyll-a estimation in productive waters is encouraging. Additionally, both models retain a band in the 400-500nm range. Bands in this range have been previously used to develop blue-green ratios for chlorophyll-a estimation (O'Reilly et al., 1998), but are prone to spectral interference from CDOM. Notable differences between Sentinel-2 and Sentinel-3 models include the retention of Rrs(740) in the Sentinel-2 model and Rrs(620)-[Rrs(709)+Rrs(560)-Rrs(709)] for the Sentinel-3 model. Bands in the near infrared range (Rrs(740)) have been used in chlorophyll-a retrieval models for turbid waters to further distinguish spectral signals of chlorophyll-a and CDOM (Pahlevan et al., 2020). The algorithm retained for the Sentinel-3 model focuses on the 620nm band, which is often used to separate phycocyanin concentrations from chlorophyll-a. Chlorophyll-a and phycocyanin are spectrally similar, however, inclusion of this algorithm in the chlorophyll-a model may indicate a limited ability of Sentinel-3 to separate the two variables in this case.

Phycocyanin models show more differences in variable selection compared to chlorophyll-a models. Variable selection for Sentinel-2 resulted in several more bands and algorithms (8) than Sentinel-3 (3) and there are few similarities between the selected variables. Both models retain the band located at 443nm and include Rrs(665) in at least one algorithm. As discussed above, both bands are often used in chlorophyll-a retrieval. In addition to Rrs(443), the Sentinel-3 model retains Rrs(412) and the Surface Algal Bloom Index developed by Alawadi (2010) to detect surface blooms in ocean environments. Notably, the 620nm band is not retained in the Sentinel-3 model, despite its established use for phycocyanin concentration monitoring. The Sentinel-3 phycocyanin model also shows relatively poor performance overall. Simis et al. (2005), found that use of the 620nm band to estimate phycocyanin resulted in high errors when the phytoplankton community was not dominated by cyanobacteria. It is possible that the relatively low concentrations of phycocyanin found in this study are difficult to capture using the 620nm band. Additionally, the datasets used here may be too small for machine learning models to capture statistical relationships between phycocyanin and Sentinel-3 reflectance. The relatively large 300m spatial resolution of Sentinel-3 imagery may also contribute to the underwhelming model performance for retrieval of all water quality variables. While Sentinel-2 does not have the 620nm band necessary to estimate phycocyanin concentrations directly, recent studies have successfully used Sentinel-2 imagery to indirectly estimate phycocyanin (R. Beck et al., 2017; Pérez-González et al., 2021; Sòria-Perpinyà et al., 2020, 2021). The Sentinel-2 models for phycocyanin and chlorophyll-a select several of the same variables, yet the selection of several different algorithms for the phycocyanin model suggests a distinct statistical relationship between band reflectance and phycocyanin compared to chlorophyll-a. Additionally, the selection of bands differencing Rrs(705) and Rrs(665) may support the algorithms developed by Wynne et al. (2008) to separate cyanobacterial blooms from other algal blooms based on spectral curve shape in Lake Erie.

The two satellite-based Pc:Chla models largely select different input variables. Only Rrs(560) is selected by both, and neither select many bands and algorithms included in the phycocyanin and chlorophyll-a models. While Pc:Chla variable selection differs between satellites, Pc:Chla models choose similar input variables to the chlorophyll and phycocyanin models for that satellite. All but one of the Sentinel-2 Pc:Chla model input variables is present in either the phycocyanin model or the chlorophyll-a model. While three of those variables are retained by all of the Sentinel-2 models, the Pc:Chla model also selects variables unique to the chlorophyll-a and phycocyanin models. This may suggest that the model is able to select variables that help distinguish between chlorophyll and phycocyanin dominated waters. The Sentinel-3 Pc:Chla model only retains three algorithms: one is unique to the chlorophyll-a model, one is unique to the phycocyanin model, and one unique to the Pc:Chla model. While the Pc:Chla models are developed independently of the chlorophyll-a and phycocyanin models, the skill of the Pc:Chla model appears to be reflected in the skill of the other algae indictor models. The Sentinel-2 Pc:Chla model generally performs well, in line with the relatively high R^2 scores for other algae indictors. The Sentinel-3 Pc:Chla model has a comparable R^2 to chlorophyll-a, and appears to perform particularly well at low phycocyanin to chlorophyll-a ratios (Figure 4-7). As Pc:Chla increases, errors generally increase. Given the poor performance of Sentinel-3 phycocyanin models, higher errors in phycocyanin dominated waters are not unexpected. While each satellite has limitations, this analysis shows that machine learning models have some ability to retrieve Pc:Chl using Sentinel-2 and Sentinel-3 imagery. Given that both green algae and cyanobacteria contain some quantity of chlorophyll-a and phycocyanin, satellite-based monitoring of this ratio, in addition to raw chlorophyll-a and phycocyanin values, may provide useful insights to lake managers regarding the characterization (green or cyanobacteria dominated) of algae blooms.

The DO model is developed independent of satellite imagery and similarly applied to algae indicators from each satellite. Differences in DO model skill between the two satellites generally follow the skill of algae models. An analysis of feature importance during cross-validation of the DO model shows that Pc:Chla contributes most to skill, with feature importance (measured by decrease in R^2) ranging between 0.5-0.9 (mean 0.67) across folds. Feature importance for both phycocyanin and chlorophyll-a range between 0.15-0.30 with average importance of 0.23 and 0.22 respectively. Given the particular importance of Pc:Chla in DO modeling, it is unsurprising that model results based on Sentinel-2 perform significantly better than Sentinel-3. The fold-averaged R^2 of DO for each satellite fall close to the skill of the respective Pc:Chla models, which may also indicate the importance of Pc:Chla model skill in the accuracy of subsequent DO model results. Recent efforts to monitor DO using Sentinel-2 imagery have often used spectral information to model DO directly (Batur & Maktav, 2018; E. A. L. Salas et al., 2022; Tian et al., 2023); research investigating modeling DO with Sentinel-3 imagery appears limited. While these existing DO models show skill locally, they are often difficult to generalize (Sagan et al., 2020). Comparatively, the model presented here is based on in-situ data and appears robust across sensors. As discussed previously, significant work exists characterizing algae conditions using algal pigments and remote sensing methods exists; therefore, given the variety of in-situ data and remotely sensed water quality models that have been made available in recent years, the DO model presented here has the potential to be applied broadly to inland waters.

4.5 Conclusions

In this study, Sentinel-2 MSI and Sentinel-3 OLCI imagery are used to develop machine learning models for the retrieval of chlorophyll-a, phycocyanin, Pc:Chla, in a small inland lake. A novel machine learning and remote sensing-based approach was also developed for the retrieval of dissolved oxygen concentrations. Machine learning models developed for both satellites showed significant capability. RF outperforms ANN in retrieval of all metrics for both satellites with the exception of modeled phycocyanin based on Sentinel-3. Overall, we find RF to be a more suitable approach for algae pigment retrieval given limited data and the ability for RF to evaluate the importance of predictors. When compared with in situ data, best model results of algae metrics based on Sentinel-2 (Sentinel-3) imagery achieved R^2 scores of 0.47 (0.42) for chlorophyll-a, 0.69 (0.22) for phycocyanin, and 0.70 (0.41) for Pc:Chla. In situ algae metric data were used to build a

RF model that indirectly estimates DO concentrations from satellite imagery, achieving an R^2 of 0.69 (0.36) when applied to Sentinel-2 (Sentinel-3) imagery. This method allows for the estimation of dissolved oxygen using algae pigment variables frequently collected for water quality monitoring in lakes. Additionally, the DO model can increase the information extracted from remotely sensed lake water quality products.

The ability of Sentinel-2 and Sentinel-3 imagery to model several harmful algae indicators on a small inland lake is encouraging from a management perspective. Spatial indication of the presence of harmful algae may be a helpful tool for the identification of blooms and to inform decision making regarding water quality testing routines, beach closings, and issuing warnings to the public. Despite lower skill for the Sentinel-3 model, the use of both satellites to retrieve chlorophyll-a, phycocyanin, Pc:Chla, and DO provides an opportunity to monitor water quality at fine spatial and temporal resolutions. Together, the use of Sentinel-2 and Sentinel-3 imagery for water quality monitoring has the potential to effectively track the development of cHABs and inform management strategies for Lake Mendota.

Chapter 5: Hydroclimatic Forecasting to Inform Anticipatory Action for Dengue Virus in Colombia

5.1 Introduction

Water – related disease is a significant contributor to morbidity and mortality, globally (Hunter, 2003). Climate and hydrology play a critical role in the emergence of vectors and pathogens. In many places, environmental constraints on vector and pathogen development result in notable inter- and intra- annual variability in vector-borne disease burden (Yuan et al., 2020). Climate and hydrology have complex and often contrasting effects on vector and pathogen biology. Increases in temperature may increase vector activity and pathogen development, but excessively high temperatures may have the opposite effect, and responses are often species and location specific (Whiting et al., 2001). The effects of precipitation are subject to even more complexity. Precipitation may increase habitat availability through ponding but may eliminate vectors in extreme precipitation events. Reduced streamflow during low precipitation periods may also provide calm waters for breeding sites (Whiting et al., 2001). Yet dry conditions may result in more outdoor household-level water storage, creating new breeding sites (Lowe et al., 2021). Given the array of contrasting hydroclimate impacts on vectors and pathogens, establishing robust relationships between hydroclimate variables and case counts remains challenging. Exploring key environmental drivers of vector and pathogen development allows for identification of conditions associated with high disease burden, which may be leveraged for the development of early warning systems.

Skillful prediction of vector borne disease incidence may improve public health preparedness, particularly on longer timescales. Recently, increasing focus has been placed on the prediction of hydroclimate conditions at timescales between short-range weather predictions and long-range seasonal outlooks. This subseasonal-to-seasonal (S2S) scale has been identified as a key timeframe for decision-making in many sectors (White et al., 2017). Public health may benefit from S2S forecasts that can inform decisions requiring longer lead times such as allocating funding, acquisition and preparation of medical supplies, or implementation of vector control strategies (Brunet et al., 2010). Exploration of forecast based early action regarding vector-borne disease may reduce morbidity, mortality, and reduce costs for government agencies and relief organizations, and may help promote forecast uptake by management agencies.

Among vector-borne diseases, dengue stands out as a global public health threat (Guzman et al., 2010). Dengue has spread significantly since its re-emergence in Latin America, with cases rising rapidly since the 1980's (Lenharo, 2023). Dengue virus (DENV) is spread to humans through the bite of infected *Aedes* mosquitoes and is considered a largely urban disease. Epidemics may become more frequent as the population of Latin America doubles by 2050 and urbanization enhances opportunities for transmission. Despite the presence of a national integrated vector control strategy, dengue transmission has not decreased in Colombia. All four DENV serotypes are actively circulating in many parts of the country and there has been a significant increase in the number of severe DENV cases since re-emergence (Gutierrez-Barbosa et al., 2020; Ocampo et al., 2014). Infection with one serotype generally provides long-term immunity, however, secondary exposure to others serotypes has been associated with increased risk of severe disease (Katzelnick et al., 2020).

Like other vector-borne diseases, incidence of DENV has been linked to climate, including local scale metrological variables that can modulate vector habitat, survival, and pathogen replication (Duarte et al., 2019; Me et al., 2018; Villegas et al., 2020). While the influence of local hydrology and climate on vector-borne disease is well documented, these variables can have interacting and contrasting effects on DENV transmission that are difficult to capture, particularly on S2S timescales. Colombia is experiencing a resurgence of vector-borne diseases, and has been identified as an emerging disease hotspot (Jones et al., 2008). Colombia also frequently experiences hydrologic extremes, often driven by large climate cycles, like the El Nino Southern Oscillation (ENSO) (Germán Poveda et al., 2011; Waylen & Poveda, 2002). Significant portions of the country are favorable to transmission of vector-borne disease (Cabrera & Selvaraj, 2020), making the region a suitable study site for the investigation of hydroclimate-vector relationships. Spatial variability in hydroclimate variables across Colombia is significant and differences in climate and land cover across Colombia may have impacts on hydroclimate-disease relationships.

This work focuses on the development of tailored statistical forecasting models for DENV at 1-, 3- and 6- month lead times for four cities across Colombia (Cali, Cucuta, Medellin, and Leticia). While several climate-based S2S warning systems have been developed for facets of DENV transmission (Muñoz et al., 2020; Tompkins et al., 2019), relatively little work has been done in Colombia. Further, in addition to public health applications, model development provides an opportunity to better understand relationships between hydroclimatic variables DENV and to assess the added value of climate information in S2S DENV forecasts. Using DENV case data from Colombia's National Public Health Surveillance System (SIVIGILA), run by the National Institutes of Health of Colombia (INS), for four cities (Cali, Cucuta, Medellin, and Leticia), S2S forecasting models are developed using local scale, and global scale hydroclimate variables. Model outputs are compared to a climatological null model and an autoregressive model at each lead time.

Specific research questions include:

1) Which local and global hydroclimate variables correlate with DENV case load at 0-, 1-, 3-, and 6- month lead times?
2) For which lead times do hydroclimate variables provide superior prediction skill compared with climatological and autocorrelation models?
3) Under which DENV conditions (high, low case load) do climate variables contribute most to prediction skill?

This modeling approach aims to better understand the dominant hydroclimatic drivers of DENV in Colombia at concurrent to seasonal lead times and leverage those relationships to develop forecast systems for season-ahead public health decision making. Forecast development provides an opportunity to investigate the role of climate information in DENV early warning systems at different S2S scales and may provide insight into the conditions under which climate-based early warning systems are valuable for DENV preparedness.

5.2 Methods

5.2.1 Study Site

Colombia has identified DENV as a significant public health threat since the 1950's. The suspension of vector control campaigns targeting *Aedes* mosquitos in 1970 led to a resurgence of dengue infections that persists today (Gutierrez-Barbosa et al., 2020). The majority of cases come from the urban areas of Colombia, driven in part by high population density and water infrastructure that may act as breeding sites for *Aedes aegypti* (Villar et al., 2015). DENV is considered hyperendemic in Colombia, due to the co-circulation of all four DENV serotypes. Several prevention measures are available to address and prevent dengue outbreaks. A dengue vaccine, Dengvaxia, is currently available. The vaccine is recommended only for people with confirmed previous dengue infection (WHO, 2019). Other effective prevention measures include the use of larvicide, insecticide (e.g., spraying, treated nets), and personal preventative measures to avoid mosquito bites (Ocampo et al., 2014; Sepulveda & Vasilieva, 2016). Despite the availability of Dengvaxia, recent work suggests that traditional vector prevention strategies in Colombia remain a viable and cost-effective option (Claypool et al., 2021). Development of tools to inform the activation of common vector control strategies in Colombia may therefore be useful for public health managers.

DENV case data is accessed from Colombia's National Public Health Surveillance System (SIVIGILA), run by the National Institutes of Health of Colombia (INS), for four cities including Cali, Cucuta, Medellin, and Leticia. Case counts are reported by clinics and hospitals to insurance agencies and regional health authorities, which are then sent to INS for consolidation. Dengue cases are defined by SIVIGILA as all people with acute febrile illness (< 7 days) with two or more of the following manifestations: headache, retro-orbital pain, myalgia, arthralgia, or rash (Rico-Mendoza et al., 2019). Case count data are available through 2021 in all cities. Data begin in 2006 in Cali and Cucuta, 2008 in Leticia, and 2009 in Medellin. Population data is available for each city through 2020, allowing for an annual estimate of DENV incidence (Figure 5-1).

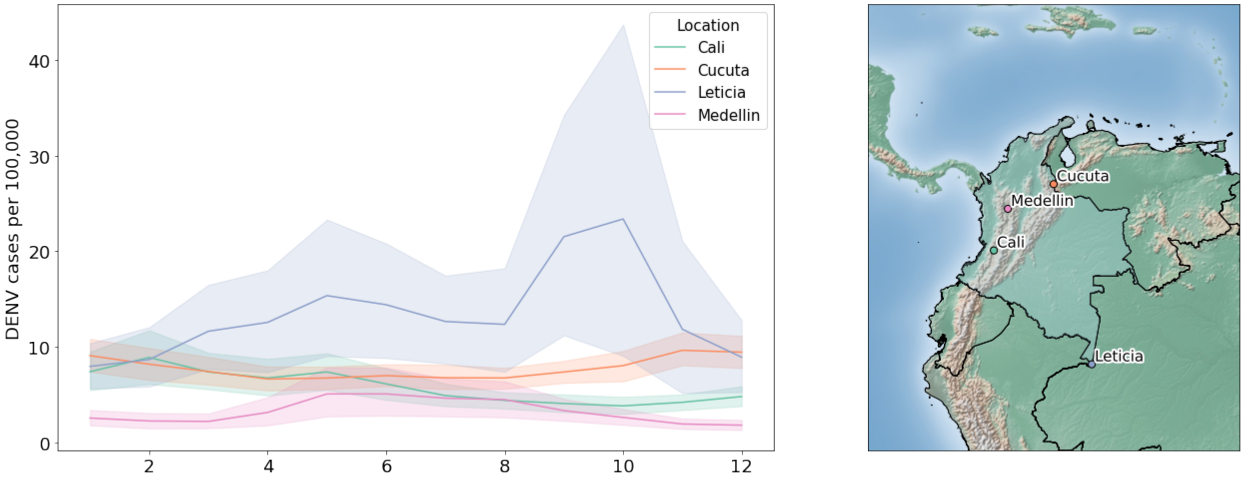

Figure 5-1. Average monthly DENV incidence per 100,000 population (i.e., climatology, left) and study site locations (right). Figure repeated from chapter 1.

Each city shows interannual variability in case load, however, DENV case seasonality is notably different among cities. Further, while each city exhibits seasons in which potential for increased DENV transmission appears higher, some years have relatively low DENV transmission across all months. Given the lack of a defined transmission season across cities, models are developed separately for each city and month, following a defined modeling structure.

5.2.2 Predictor Selection

Predictors are evaluated in a correlation analysis at 0-, 1-, 3-, and 6- month lead times. Correlations are calculated separately for each city and each month of the year (i.e., Cali January DENV incidence is correlated with Cali January Temperature for the years available). Candidate predictors are chosen based on a literature review. Climate and hydrologic variables are considered at local and global scales. Correlations are evaluated at concurrent (0-month lead) timescales to assess whether forecasts of hydroclimate variables from the North American Multi-Model Ensemble (NMME) may be useful in predictions of DENV incidence. Seven variables are evaluated based on their theoretical relationship to DENV incidence (Table 5-1).

Table 5-1 Variables used in correlation analysis with monthly DENV incidence at 0-, 1-, 3- and 6- month lead times.

Potential Predictors (monthly)	**Source**	**Resolution**
Total Precipitation [mm]	IDEAM	Gauged
Relative Humidity (mean, max, min) [%]	IDEAM	Gauged
Temperature (mean, max, min) [°C]	IDEAM	Gauged
Streamflow (Leticia only) [m^3/s]	DHIME – IDEAM	Gauged
ENSO Regions (1+2, 3, 3.4, 4) [°C]	NOAA ERSST v5	2°
Global Sea Surface Temperature [°C]	NOAA ERSST v5	2°
Global Geopotential Height (200 mb) [gpm]	NCEP-NCAR	2°

As discussed above, temperature and precipitation have been widely shown to influence pathogen and vector development, and dengue transmission at lead times from weeks to months (Johansson et al., 2009). Additionally, hydroclimatic extremes are strongly influenced by large-scale climate phenomena at S2S scales. Such phenomena often develop slowly and may provide prospects for predicting the likelihood of hydroclimatic extremes and DENV outbreaks from weeks to months in advance.

Phenomena such as ENSO often modulate local climate conditions (e.g., temperature, humidity, and precipitation) through atmospheric teleconnections (Barnston, 1994; Giannini et al., 2000; Markowski & North, 2003). Additionally, Leticia is located on the banks of the Amazon River, thus streamflow is also considered as a candidate predictor, given the potential for the river to create vector habitat under certain flow conditions. The effects of ENSO on the climate and hydrology of Colombia are well-established and influential (Germán Poveda et al., 2011). The climatic influence of ENSO has previously been linked directly to the incidence of vector-borne disease in Colombia (Germain Poveda et al., 2000). Given this established connection, indices for ENSO regions 1+2, 3, 3.4, and 4 are evaluated as candidate predictors of DENV incidence. These one-dimensional (i.e., gauged data and climate indices) predictors are evaluated in a simple correlation analysis with DENV incidence at each of the relevant lead times (Figure 5-2).

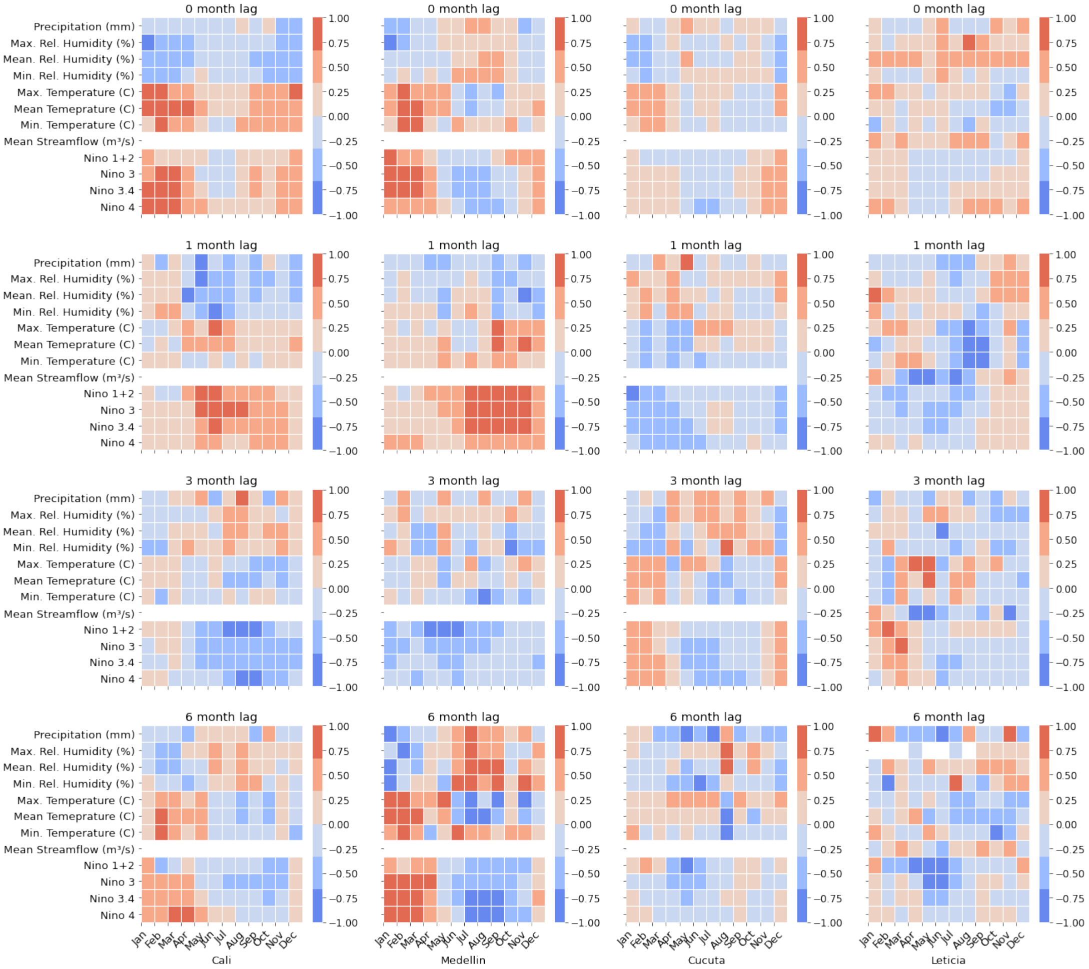

Figure 5-2. Correlation analysis results for one dimensional candidate predictor variables

A significant potential source of error when correlating gauged local scale predictors and DENV incidence is the unknown size of the case reporting area. Case data from SIVIGILA is collected through passive surveillance at healthcare facilities. This creates a potential mismatch between the scale of local hydroclimate data measured at a point, and the reporting region for DENV cases which could be significantly larger. To combat this source of uncertainty, a sensitivity analysis is performed in which correlations between DENV incidence and gridded temperature, and precipitation (Funk et al., 2015) data are assessed within the municipal district surrounding each city.

Colombia's climate and hydrology are also influenced by a variety of low level jets with origins in the Caribbean and coastal tropical pacific (H. D. Salas et al., 2020). To capture the variable sources of large-scale hydroclimate influence, statistically derived regions of global sea surface temperature (SST) data are evaluated as potential predictors. Similarly, regions of geopotential height are known to play a significant role in atmospheric circulation patterns and may influence the distribution of weather across South America. One such feature is the Bolivian high, an upper-level anticyclone that manifests as a region of high pressure. The Bolivian high is thought to have significant impacts on temperature and precipitation across South America (Lenters & Cook, 1997). Therefore, statistically derived regions of 200 mb geopotential height are used as candidate predictors.

Regions of relevant global predictors are identified in a Monte Carlo analysis to ensure a higher level of confidence in the statistical relationship between climate variables and DENV incidence. Correlations between DENV incidence and each climate variables are calculated globally. The number of significant correlations is then compared to 1000 trials in which the target variable is randomly shuffled (following Zimmerman et al., 2016). This process helps to ensure that spurious correlations are removed, ultimately providing a more robust selection of global climate predictors. This selection process is performed for each city in each month. If any regions are selected, a principal component analysis is performed on the climate variable of interest in those regions. Principal components that explain more than 10% of the variance in DENV case counts in regions of interest are then evaluated as candidate predictors.

5.2.2 Modeling Approach

Model hindcasts for DENV are developed in selected cities for 1-, 3-, and 6- month lead times. Given the variability in seasonality of DENV incidence across cities, models are developed separately for each month. For example, to forecast January DENV incidence in Cali, a model is developed with data from 1-, 3- and 6- months prior. Following this structure models are developed for each month. This approach allows for evaluation of model skill, and the statistical power of climate information, in each month.

A random forest regression model is chosen to model DENV incidence and is implemented using the scikit-learn package in python (Pedregosa et al., 2011). Random forests are a non-parametric ensemble learning method based on the construction of many decision trees, each collectively voting on an outcome (Breiman, 2001). Random forests are well-suited to the task of modeling DENV incidence given the non-normal distribution of the data. Additionally, random forests are more robust to overfitting compared to deep learning methods. This is a particularly important feature given the limited DENV incidence time series. Random forests also have the benefit of computing variable importance (marginal decrease in model performance when each variable is excluded) during the modeling process. This allows for a direct evaluation of climate variable importance among cities, lead times, and months. A leave one out cross validation (LOOCV) hindcast approach is adopted to evaluate model performance. For each city and month specific model, one year of the timeseries is dropped, the model is constructed, and the missing value is predicted. This approach is conducted iteratively until all years have been predicted, allowing for an estimation of model skill. Uncertainty in model outputs is estimated using the forestci package, which implements a monte-carlo based estimation of variance for random forest regression (Wager et al., 2014).

An autoregressive null model is constructed to compare against each random forest model. Autoregression is commonly used for disease transmission forecasting (Baharom et al., 2022). An autoregressive model is constructed for 1-, 3-, and 6- month lead times for each month in each city using linear regression. A null model of DENV incidence is also evaluated against random forest and autoregression models, defined as the long term (using all available data) average incidence for each month in each city (Figure 1). This historical average is what might be predicted in an environment where no other information is available about the system.

5.2.3 Model Evaluation

Models are evaluated deterministically and categorically. The coefficient of determination (R^2) is used to evaluate deterministic model performance. Accuracy, sensitivity, and specificity are evaluated to determine categorical performance.

Dengue incidence categories are defined using the endemic channel. The endemic channel is a tool to estimate the central tendency, along with upper and lower limits of epidemiological data first described by Bortman (1999). The endemic channel calculation used here is adapted from the R package epiCo, a software package developed specifically for evaluation of vector borne disease in Colombia (Umaña et al., 2024). An endemic channel is calculated for each city at a monthly interval. The central tendency is calculated for each month by taking the geometric mean of historical case data for each month. The lower and upper limit are then calculated using the geometric standard deviation, representing a 95% confidence interval. Epidemic years are commonly removed from the calculation of the endemic channel to provide a better estimate of 'normal' disease conditions. Here, monthly case counts are omitted from the calculation if they are greater than the 90th percentile of observed cases in each city. The deterministic model results are binned into the four categories represented by the endemic channel: *Below Safety, Above Safety, Warning,* and *Epidemic*. These categories are then used to evaluate accuracy, sensitivity, and specificity of each model at 1-, 3- and 6- month lead times.

Variable importance is also assessed for each model. Variable importance is calculated as the mean decrease impurity, calculated for each feature as the total decrease in node impurity weighted by the probability of reaching that node, averaged over all trees. Variable importance is assessed for each model. For comparison, variable importance is then grouped into categories based on variable type and source to better understand the influence of different processes (Table 5-2).

Table 5-2 Predictor categories for variable importance assessment.

Source Category	Type Category	Predictor
Autocorrelation	Autocorrelation	DENV Incidence
Lagged	Hydrology	Total Precipitation [mm]
Lagged	Hydrology	Relative Humidity (mean, max, min) [%]
Lagged	Hydrology	Streamflow (Leticia only) [m^3/s]
Lagged	Temperature	Temperature (mean, max, min) [°C]
Lagged	Global Climate	ENSO Regions (1+2, 3, 3.4, 4) [°C]
Lagged	Global Climate	Global Sea Surface Temperature [°C]
Lagged	Global Climate	Global Geopotential Height (200 mb) [gpm]
NMME	Temperature	GFDL SPEAR Total Precipitation [mm]
NMME	Temperature	GFDL SPEAR Temperature (max, min) [mm]
NMME	Global Climate	GFDL SPEAR ENSO Regions (1+2, 3, 3.4, 4) [°C]

The LOOCV process creates a model structure corresponding to each month of DENV incidence data. Variable importance is calculated for each model realization in the LOOCV process. Variable importance can then be compared among lead times, cities, months, and endemic channel categories to answer questions regarding the role of climate information in predicting DENV incidence under a wide variety of conditions.

5.3 Results

5.3.1 Leading Predictors

Predictor relevance is evaluated two ways: by the number of times a predictor is retained for a model (expressed as a percent), and the mean decrease in impurity (variable importance) calculated for a predictor when it is included in the random forest model structure.

The three most commonly retained predictors across all cities, months, and lead times, with an importance score greater than zero are autocorrelation, the first principal component of sea surface temperatures, and lagged Nino regions (3 and 3.4) (Table 5-3). The similarity in the most commonly used predictors indicates the outsized importance of both autocorrelation and large-scale climate features in prediction of DENV incidence. Notably, the percent inclusion and average importance vary for these predictors among lead times. At lead times of 1-month, autocorrelation dominates. One month lagged DENV incidence is included as a predictor in 100% of models and on average accounts for 56% of the mean decrease in impurity (variable importance). While SST PC1 and Nino 3 SSTs (one month lag) are the next most important variables, they appear in only 35% and 23% of models, and have an average importance of less than 10%. As expected, the influence of autocorrelation declines in the 3- and 6- month lead times, included in only 69% and 27% of models, respectively. While the percent inclusion of sea surface temperature and Nino predictors remains relatively similar at 3- and 6- month leads, average variable importance increases notably for both, as importance of autocorrelation declines. This generally indicates a shift in predictive power from autoregressive features to large scale climate features as lead time increases.

By type, the most prevalent predictors largely reflect the patterns indicated by the top individual predictors. At 1-month lead time, 100% of models include autoregressive features and 70.8% include global climate, followed by temperature predictors and hydrology predictors. At the 3-month lead time percent inclusion of autoregressive and global climate predictors in models is the same. At the 6-month lead time, inclusion of global climate predictors leads autoregressive predictors 67% to 27%. Temperature and hydrology predictors are ranked third and fourth, respectively, at all lead times.

Table 5-3 Three most retained predictors for each lead time, across all cities and months with importance scores greater than zero.

Lead Time	Predictor	Inclusion (N=48)	Average Importance
1-month	Autocorrelation	100%	56%
	Sea Surface Temperature PC1	35.4%	8.4%
	Nino 3.4 (1-mo lag)	22.9%	2.6%
3-month	Autocorrelation	68.8%	38.1%
	Sea Surface Temperature PC1	37.5%	15.1%
	Nino 3 (3-mo lag)	25%	2.6%
6-month	Sea Surface Temperature PC1	35.4%	30.1%
	Autocorrelation	27.1%	19.7%
	Nino 3.4 (6-mo lag)	21%	7.8%

Predictor inclusion patterns are generally similar to overall results when broken down by city, with some notable exceptions. In all cities, autoregressive features decline in percent inclusion and global climate features tend to increase with lead times. Local temperature and hydrologic predictors are generally included in fewer models. Models in Leticia use hydrologic predictors more frequently than other cities. This is largely driven by the inclusion of Amazon streamflow as a predictor in Leticia, which is retained for several months of the year at 1- and 3- month time lags.

Higher relative inclusion of temperature predictors is seen in Cali and Medellin models, compared to Leticia and Cucuta. These differences are largest at the 1- and 3- month lead times. This difference is driven by the inclusion of NMME forecasts of maximum and minimum temperature in these cities. DENV incidence in Cali and Medellin is strongly correlated with temperature metrics at concurrent lead times during the beginning of the year. These relationships are weaker in Leticia and Cucuta. The absence of temperature indices at the 6-month lead might indicate a decrease in NMME temperature forecast skill at longer lead times.

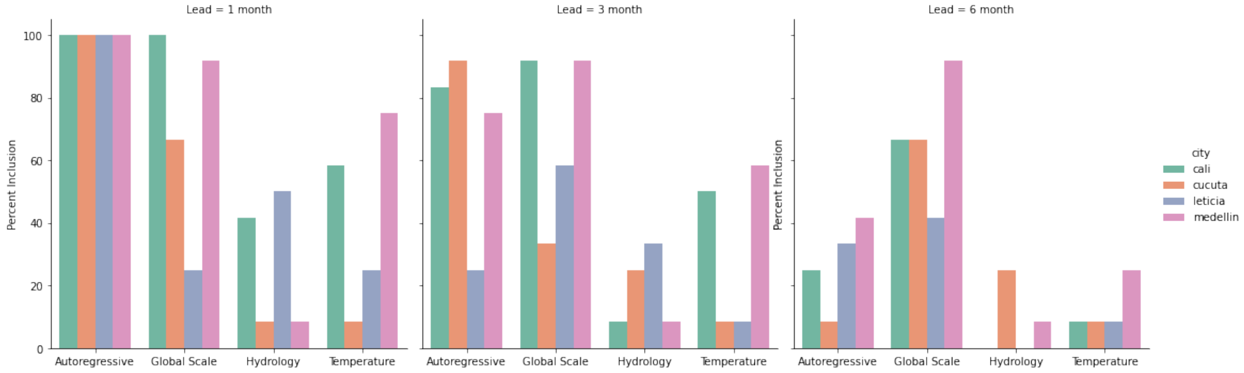

Figure 5-3. Percent inclusion of predictor types by city and lead time.

5.3.2 Model Performance

At the 1-month lead time, 48 (N=48) city and month -specific models have at least one significantly correlating predictor with which to construct a random forest model. At 3-month and 6-month leads, 47 and 38 models are constructed, respectively. Months without a significantly correlating predictor default to climatology (the long-term average of DENV cases, specific to each city and month). Model skill is evaluated in a point-by-point comparison (deterministic) and a categorical comparison (categorically).

R^2 values are calculated for each city and month specific model and are compared to an autoregressive model with a climatological null as the baseline. The climatological null model always has an R^2 of zero. In some cases, autoregressive models are particularly poor resulting in negative R^2 values. In these cases, comparing random forest performance against autoregressive models can inflate the gain in skill seen by the random forest. Therefore, the R^2 values for the null models range from zero to one.

At the 1-month lead only 6 (12.5%) models make improvements over autoregressive models and climatology (Figure 4). 25 (52%) models improve over null models at the 3-month lead time and 20 (42%) models improve over null models at the 6-month lead. For random forest models that showed improvement over null comparisons, the average increase in R^2 of random forest models over autoregressive models and climatological models (whichever R^2 is higher) is 0.2 at the 1-month lead, 0.24 at the 3-month lead, and 0.32 at the 6-month lead.

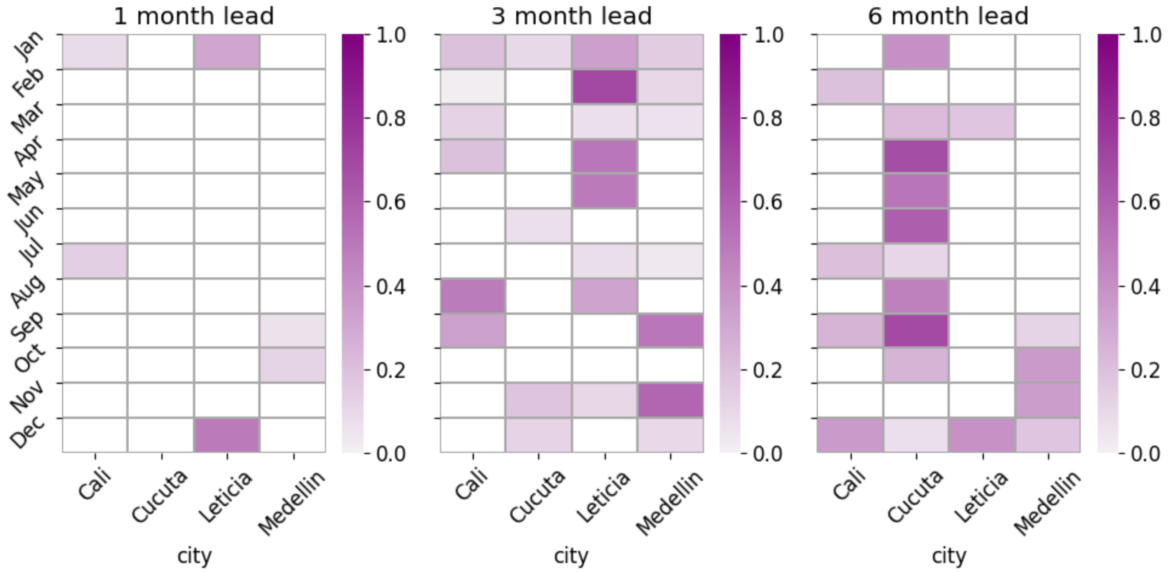

Figure 5-4. Improvements in random forest R^2 (coefficient of determination) over autoregressive and climatological null models.

On average, improvements in forecast skill are greatest at the 6-month lead. These improvements, however, are concentrated in Cucuta, where 10 of the 12 month-specific models outperform autoregressive and climatological null models. Improvements are present in other cities but are more sporadic across the year. While gains at the 3-month lead are more modest, performance over the null models appears more consistent, particularly at the beginning and end of the year. This might indicate the presence of a more reliable climatic influence on DENV transmission conditions at the 3-month lead across cities.

Forecast accuracy, sensitivity, and specificity are also calculated for each city, lead time, and category derived from the endemic channel. Accuracy represents the proportion of predicted categories that match observed categories. Sensitivity (true positive rate) measures how often a category is correctly predicted, conditioned on the total number of times that category was observed (e.g., number of *epidemic* months predicted relative to the total number of *epidemic* months observed). A forecast with perfect sensitivity will always capture the occurrence of the category of interest (minimizes false negatives). Specificity (true negative rate) measures how often a category is correctly discarded conditioned on the total number of times that category was not observed (e.g., number of non-*epidemic* months predicted relative to the total number of non-*epidemic* months). A forecast with perfect specificity will never predict the category of interest when it is not observed (minimizes false positives). Each metric is measured from 0 to 1.

For random forest models, all categorical metrics tend to decline from short to long lead times. Accuracy and sensitivity of forecasts are largely stable among categories and lead times. Accuracy ranges from 0.69 to 0.85, and specificity ranges from 0.72 to 0.88 across all categories and lead times. Accuracy and specificity of the *epidemic* category experience the largest declines from 1- to 6-month leads, compared to other categories. This suggests that the *epidemic* category is particularly susceptible to false positives at longer lead times, or overpredictions of DENV incidence. Sensitivity has the greatest variability among categorical metrics, ranging from 0.17 to 0.79. On average, sensitivity is lowest in the *above safety* and *warning* categories. This implies that models are often predicting different categories when the observed category is between the two extremes (*below safety, epidemic*). Sensitivity is highest for the *epidemic* category, which is encouraging given that public health officials are likely most concerned with accurate predictions of high transmission scenarios.

Compared to predictions made by the autoregressive null model, the inclusion of climate information in the random forest model appears to make some improvements. On average, random forest shows slight improvement in sensitivity (0.02), accuracy (0.04) and specificity (0.02) over the autoregressive model. Nearly all increases in categorical skill come from the 3-

and 6- month lead times. The greatest improvements in accuracy and specificity are seen in in 3- and 6- month leads of the *epidemic* category (Figure 5-5). Sensitivity shows the most improvement at the 3- and 6- month leads of the *below safety* category. This may imply that the addition of climate information improves detection of months in which transmission will be low, compared to models based solely on previous case data. The number of false predictions for epidemic months is also reduced, leading to improved specificity. This is largely driven by a decrease in the number of observed *below safety* events predicted to be *epidemic* events. It is also important to note that the inclusion of climate information sometimes results in a decrease in categorical skill compared to a simple autoregressive model. Declines are likely related to the choice of random forest for model construction, compared to linear regression for the autoregressive models.

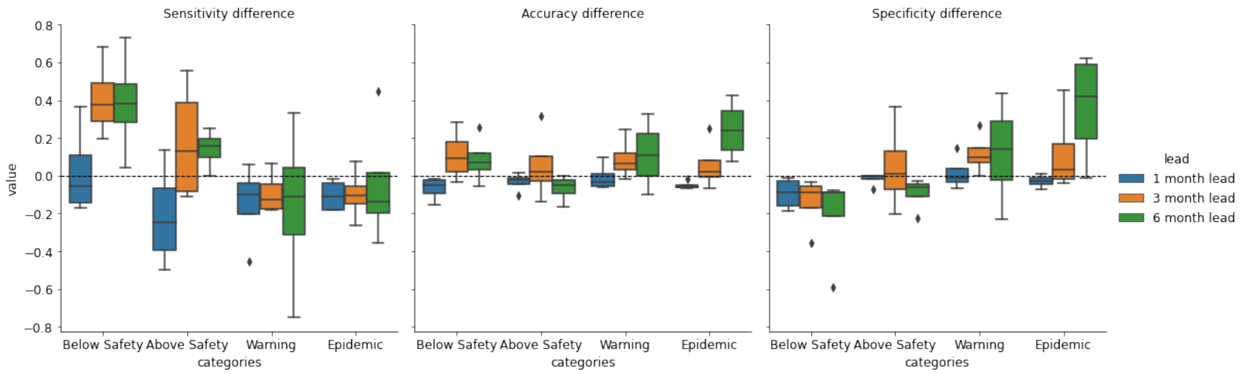

Figure 5-5. Difference between random forest and autoregressive model forecast sensitivity, accuracy, and specificity.

5.3.3 Variable Importance

To evaluate the transmission conditions under which climate information is most useful for DENV incidence prediction, feature variable importance is assessed for all random forest models (Figure 5-6). Feature importance is compared across a gradient of transmission conditions. To compare low transmission conditions to high transmission conditions, the boundary indicating the start of the *epidemic* category in each month and city is ranked from lowest (1) to highest (48). This provides a proxy for average transmission conditions in each month and city. Feature importance is compared average across these ranks.

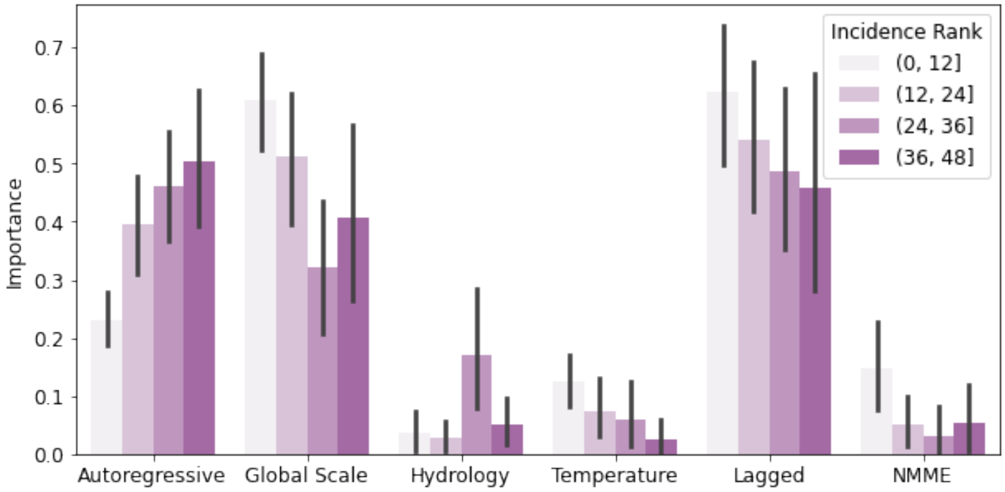

Figure 5-6. Average feature importance for predictor categories (see Table 2), binned by incidence rank. Low values correspond to low average DENV incidence and vice versa.

Autoregressive features are found to increase in importance from low transmission scenarios to high transmission scenarios. Conversely, nearly all categories of hydroclimate predictors decrease in importance from low to high transmission scenarios. This suggests that hydroclimate predictors are more effective in environments where DENV incidence tends to be lower based on long-term averages and autoregressive predictors perform better when DENV incidence is expected to be higher. This finding is also reflected by the months and locations in which random forest models outperform null models. Skillful city and month specific random forest models (based on R^2) have a median incidence rank of 19. Models that were outperformed by autoregressive and climatological null models have a median incidence rank of 26. Categorical performance of random forest models also indicates that climate information improves skill for prediction of low DENV incidence, given the increases in sensitivity for the *below safety* category and increases in specificity for *epidemic* categories, particularly at longer leads.

5.4 Discussion

5.4.1 Lead Time Characteristics

Development of climate-based forecasts for DENV incidence at multiple lead times provides insight into the potential for skillful prediction of DENV conditions and the predictors driving skill at varying forecast leads. Of the city and month specific random forest models that improve upon null models, a sizeable majority are at 3- and 6- month lead times. This suggests limited ability of climate variables to improve upon the predictive signal of autoregressive features at the 1-month lead. For the six random forest models that did outperform null models at the one-month lead, all still had high degrees of autocorrelation (>0.8). The two models with the largest improvement at the 1-month lead were in December and January at Leticia. These models both had high degrees of autocorrelation, but slightly less than average at the 1-month lead. Additionally, each model had at least one highly correlating climate-based predictor. This may suggest that climate variables must have a strong relationship with DENV incidence at the one-month lead to improve on autoregressive models.

Given the strong performance of autoregressive models at the 1-month lead, climate-based S2S DENV forecasting efforts may be better suited to longer lead times. At 3- and 6- month lead several autoregressive models perform worse than climatology. In these scenarios, climate-based models are particularly well positioned to make meaningful improvements in season-ahead DENV prediction. At the 3- and 6-month leads, autoregressive models perform worse than the climatological null in 27 instances. In these instances, random forest models conditioned entirely on climate variables show notable skill in DENV prediction, with average R^2 of 0.33 (0.03 – 0.72) and 0.32 (0.07 – 0.69). In Cucuta, skillful autoregressive models are found in three months. Skillful climate models are constructed for ten months out of the year, improving on null model R^2 by a margin of 0.4, on average. This analysis shows that longer seasonal lead times, in locations with low autocorrelation among cases have the potential to benefit significantly from the development of climate based DENV forecasting models.

Forecast development at multiple lead times also provides a unique opportunity to assess the performance of predictors at different time scales. As expected, the performance of autoregressive predictors tends to decline as lead times increase. Global climate predictors, including ENSO regions, regions of global SST, and regions of geopotential height, show the largest increases in importance and retention as lead times increase. These increases are driven largely by lagged regions of global SSTs. Nino indices also contribute to this increase, both through the inclusion of lagged Nino indices and GFDL SPEAR forecast outputs. This is particularly pronounced at the 6-month lead time when importance of autoregressive features is significantly lower. The memory of sea surface temperature predictors is also expected to be longer than many local climate variables, which may contribute to their increased importance at longer lead times.

Statistically derived regions of SSTs tend to have higher average variable importance than Nino indices. One possible reason for this is that statistically derived regions already capture much of the ENSO signal. For example, at the 6-month lead time in Medellin many of the selected SST regions fall within one of the defined Nino regions (Figure 5-7). This is reflected in correlations with Nino indices, where January – April DENV incidence is highly correlated with Nino indices at a 6-month lag (July-October) (Figure 5-2). In addition to the ESNO signal, statistically derived SSTs may capture additional regions of global teleconnection. This may increase the number of months in which SSTs have a measurable relationship with DENV incidence, increasing variable importance.

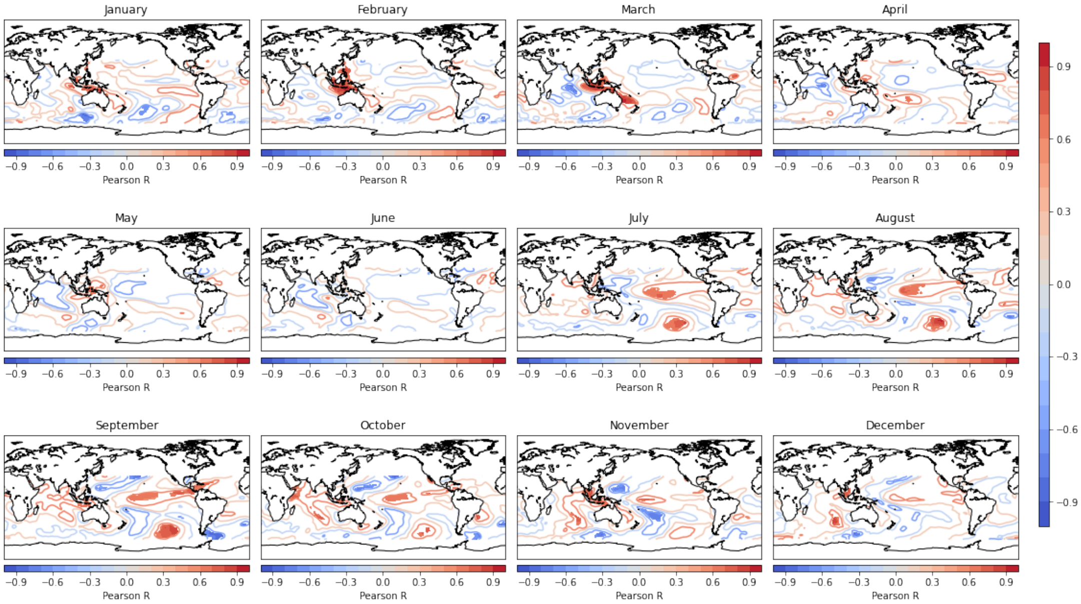

Figure 5-7. Correlations between SSTs and Medellin DENV incidence 6 months ahead. Correlations are indicated by contour lines. Regions selected as candidate predictors are filled.

Like Nino indices, temperature (min, mean, and max) variables show strong correlation with DENV incidence, but tend to appear in models less frequently as lead time increases. This may be partially driven by decreasing skill in GFDL SPEAR temperature forecasts at longer lead times, compared to forecasts of Nino indices (sea surface temperatures). Hydrologic predictors are also included in fewer models as lead times increase, and importance is low overall.

5.4.2 Dengue Conditions

In addition to assessment of lead time performance, development of DENV forecasts may benefit from a better understanding of the conditions under which climate-based forecasts perform best. Categorical performance indicates that the addition of climate information tends to improve accurate forecasts of low DENV incidence categories. Comparing categorical results between random forest and autoregressive models reveals large increases in the number of *below safety* events that are correctly predicted at the 3- and 6- month leads, with an 84% and 100% (6-month autoregressive has no accurate *below* safety predictions) increase at each lead, respectively. These gains appear to come from better discrimination of extremely high and low dengue conditions at the 3- and 6- month lead times (Figure 5-5). This is most apparent in the number of *below safety* months categorized as *epidemic* by each of the models. At the 3- month and 6- month leads, these mis-categorizations decline by 42% and 43%. This points to a weakness of autoregressive models in capturing low DENV transmission years.

An analysis of variable importance reinforces these results. When comparing average DENV conditions (as incidence rank) to variable importance, it becomes clear that climate variables are more important in the cities and months in which DENV incidence is typically low, and autoregressive features are more important when DENV incidence is high (Figure 5-6). This also suggests that climate-based models perform best in scenarios where average DENV incidence is low. To verify this finding, incidence rank is compared to the percent of categorical predictions corrected by the random forest model over the autoregressive model. Using the same groupings for incidence rank discussed above, at the 3- and 6- month lead, improvements in categorical forecast performance are concentrated in the cities and months with low average DENV caseload, largely driven by better performance in the *below safety* category. For both leads, some notable improvements are also seen in both the cities and months with highest average DENV case load, driven by a reduction in the number of *below safety* events categorized as *epidemic* events. Further, strong performance of autoregressive features in months and locations typically experiencing high dengue conditions may be the result of better data availability regarding dengue outbreaks. Increases in dengue incidence in previous months have well-established and robust relationships with the occurrence of dengue outbreaks in these scenarios. Comparatively, scenarios experiencing more infrequent outbreaks must rely more heavily on climate conditions to indicate the suitability of dengue transmission.

Together, the results of this modeling effort suggest that integration of climate information into DENV forecasts is particularly effective for accurate prediction of low DENV incidence at 3- and 6- month lead times. This feature is seen most frequently in the cities and months that often experience low incidence; however, some benefit is also seen in locations with high average DENV incidence, due to correction of extreme categorical misses. This may be related to relatively poor autocorrelation in DENV incidence at the 3- and 6- month lead times. While some improvements are seen in prediction of *epidemic* conditions, they are more modest.

5.5 Conclusions

In this chapter, tailored statistical forecasting models for DENV incidence are developed at 1-, 3- and 6- month lead times for each month in four cities across Colombia (Cali, Cucuta, Medellin, and Leticia) to inform public health decision making at seasonal lead times. Forecasting models are constructed using a random forest structure and are conditioned on autoregressive and hydroclimatic variables. A purely autoregressive model is compared as a null. Hydroclimate variables include global scale features such as sea surface temperatures and geopotential height, as well as local scale temperature, precipitation, and relative humidity. Development of forecasts provides an opportunity to evaluate the relative contribution of hydroclimate variables to DENV predictions, lead times at which climate-based forecasts perform best, and the conditions under which DENV forecasts perform best.

At the 1-month lead, autocorrelation is the dominant source of predictive skill, included in 100% of models, and is rarely outperformed by climate variables. Autoregressive features are included in fewer models at the 3-month and 6-month leads. At these longer lead times, global regions of sea surface temperature and Nino indices are most predictive of DENV incidence, followed by temperature. These predictors indicate the importance of temperature range in the development of mosquitos, pathogens, and contact with people.

Given the increased importance of hydroclimate predictors at the 3- and 6-month lead times, forecasts are found to outperform autoregressive null models more frequently at these lead times. 25 (52%) models improve over null models at the 3-month lead time and 20 (42%) models improve over null models at the 6-month lead. The magnitude of improvement is greatest in city and month specific models where autoregressive features have no predictive skill. Categorical accuracy ranges from 0.69 to 0.85, and specificity ranges from 0.72 to 0.88 across all categories and lead times. Sensitivity has the greatest variability among categorical metrics, ranging from 0.17 to 0.79. Improvements in categorical metrics over the null model are concentrated in the *below safety* and *epidemic* categories, driven by reduction in mis-categorizations of *below safety* as *epidemic*. Given these results, climate-based models appear to be most effective at 3- and 6-month leads, in scenarios where average DENV incidence is low and autocorrelation shows little predictive skill.

Development of climate based DENV incidence forecasts appear to have potential to better inform public health interventions, particularly at longer seasonal lead times. Careful consideration of the strengths of such forecasts under different climatic and public health scenarios may help target dengue forecast development to improve public health preparedness and response.

Chapter 6. Summary and Conclusions

Hydroclimatic variability exerts significant influence over the function of natural and human systems. In addition to the immediate risks associated with hazards, biological outcomes of concern are often influenced by hydroclimate variability and extremes. Two prominent water-related outcomes of concern are harmful algae and dengue virus. Understanding the relationships between hydroclimate variability and these biological outcomes can aid in the development of novel tools to improve disaster preparedness and inform water resources management. This dissertation explores the relationship between hydroclimate variability, water quality, and water related disease, and aims to advance the development of hydroclimate and satellite-based tools to improve water resources management and public health decision-making related to these biological outcomes. Specifically, seasonal forecasts for harmful algae are developed and translated to lakes across the U.S., a satellite-based harmful algae and dissolved oxygen monitoring tool is built for a small inland lake, and seasonal forecasts of dengue virus incidence are created and evaluated at multiple lead times and in several cities across Colombia. While this work is pursued through case studies, results illustrate the efficacy of hydroclimate-based season ahead forecasts for biological applications and provide insight into the relationships between climate, hydrology, and biological outcomes that may be widely applicable.

6.1 Objective one: Develop and assess targeted season-ahead forecasts of biological outcomes conditioned on hydroclimate variables for use in management of water quality and water-related disease.

Chapters 2 and 3 address the research questions *Can hydroclimate information be used to skillfully predict harmful algae outcomes and beach closings at seasonal lead times in small inland lakes? (Chapter 2)* and *are hydroclimate-based seasonal harmful algae forecasts transferable to other lakes at scale, and what lake characteristics are associated with skillful models (Chapter 3)?* A statistically-developed lake-specific cyanobacteria and beach closing forecast model in Lake Mendota demonstrates significant skill in prediction of peak season (June-August) cyanobacteria biomass at 2- and 3- month lead times. Integration of global and local hydroclimate information make significant contributions to model performance in Lake Mendota, particularly for conditons of high cyanobacteria abundance. Extreme precipitation events and seasonal streamflow are shown to be important at local scales. Additionally, teleconnections between large-scale climate phenomena and local hydroclimate conditions are shown to influence cyanobacteria abundance on seasonal timescales. These results highlight differences in the relative importance of unique drivers of cyanobacteria biomass in different mean states of the atmospheric/oceanic system. Extrapolation of this modeling framework to lakes across the northeast and Midwest U.S. reveals that integration of climate information into season-ahead forecasts of harmful algae results in skillful prediction of algae abundance and persistence across peak seasons for the majority of study lakes. Further, this modeling effort affirms the previous finding that extreme precipitation events and global sea surface temperature teleconnections play a role in shaping algae productivity in a wide variety of lakes. Pre-season algae abundance is also found to be a strong predictor of peak season algae conditions. Finally, this large-scale modeling effort suggests that static lake characteristics play a role in forecast skill.

Chapter 5 applies the tools developed across earlier chapters to forecasting dengue virus in Colombia. Specifically, chapter 5 addresses the research question *can hydroclimate information be used to skillfully predict dengue virus incidence, and if so, in what scenarios is hydroclimate information useful (Chapter 5)?* The integration of climate information into seasonal forecasts of dengue virus is promising. Dengue virus expresses unique seasonality across the four study sites in Colombia, thus tailored forecasting models are developed for each city and month. Autoregressive models, the common default approach, provide a baseline for comparison of forecast skill; these models perform very well at one month lead times but fail to provide meaningful information at longer leads when the importance of climate information in dengue forecasts increases. Most notable is the importance of large-scale climate phenomena, including the El Nino Southern Oscillation, in contributing to forecast skill. These global climate indices partially modulate temperature across Colombia, which in turn constraints the development of vectors and pathogens. The persistence and memory of certain global climate features allows dengue virus forecasts at long lead times, up to 6 months or more. Forecast skill is also linked to the average dengue case load in each city and month. Climate-based forecasts are particularly adept at categorically separating between extremely low and high dengue incidence. Careful consideration of local conditions and the strengths of climate-based forecasts may help target forecast development to provide public health information at longer leads.

While the applications and study sites of chapters 2, 3, and 5 vary, the common modeling framework allows key themes to emerge:

1) Water-related biological outcomes can be modulated by hydroclimate conditions at seasonal timescales, leading to forecast development and potentially informing water resources management.
2) The state of the atmospheric/oceanic system influences local-scale drivers and water-related biological outcomes.
3) Physical characteristics of water systems often control the variability in biological outcomes explained by hydrology and climate.

Future related research could explore optimal thresholds to trigger actions for public health interventions related to harmful algae and dengue virus. While forecasts are skillful, the actions taken based on forecasted conditions are still loosely defined in both case studies. Future work with decision-makers in Madison, Wisconsin, and Colombia can help operationalize these products.

Additionally, questions remain regarding the occurrence of hydroclimate extremes and their relationships to both harmful algae and dengue virus at seasonal timescales. Future work aimed at better understanding the links between global climate, local extremes, and biological outcomes at seasonal timescales will serve to improve forecast skill and may provide insights into the co-occurrence of biological and climate-based disasters.

6.2 Objective two: Explore the ability of satellite remote sensing to retrieve water quality metrics and differentiate between harmful algae indicators to improve monitoring capabilities.

Chapter 4 addresses the research question *can satellite imagery and machine learning methods improve monitoring and discrimination of algae on a small inland lake (Chapter 4)?* This chapter focuses primarily on monitoring strategies for water quality outcomes at weekly lead times as opposed to seasonal timescales. Satellite based monitoring tools provide a valuable supplement to in-situ water quality sampling and allows for spatial representation of harmful algae and other water quality dynamics. The wide array of harmful algae monitoring algorithms for Sentinel-2 and Sentinel-3 satellite missions are used in a machine learning structure to retrieve indicators of green algae, cyanobacteria, and dissolved oxygen. Machine learning models demonstrate skill in retrieval of algae indicators for both satellites. Despite the finer spectral resolution of Sentinel-3 imagery, the model conditioned on Sentinel-2 imagery outperformed the Sentinel-3 model for all variables. The fine spatial resolution of Sentinel-2 may offer a better estimate of reflectance values where in situ data were collected compared to Sentinel-3, particularly if algae conditions are variable across the lake. Indirect, satellite-based estimates of dissolved oxygen are modeled using machine learning methods, with algae pigments as predictors. Spatially, model errors appear to be higher when algae pigment and dissolved oxygen variability is higher. Further research could explore the drivers of spatial variability in model error, and how they may be addressed. Additionally, the use of satellite-based algae monitoring has the potential to inform decision-making regarding water quality testing and beach closings on Lake Mendota. Further work on testing and operationalizing this tool for Lake Mendota is warranted.

6.3 Final words

The use of hydroclimate and satellite information for monitoring and forecasting biological outcomes in water systems shows promise for improving water resource management. Often, water quality and water-related disease are viewed as a step removed from hydrologic and climate systems. Connecting global and local climate to ecosystem dynamics improves our understanding of the relationship between hydroclimate extremes and biological outcomes of concern and can foster the development of tools to better prepare for public health emergencies at seasonal leads. Combining forecasts with satellite monitoring effectively addresses biological water resource management challenges at multiple temporal and spatial scales. Further exploration is necessary to appropriately tailor applications of these research outcomes for water resources decision-making and public health risk reduction. A future defined by climate uncertainty warrants continued research at the intersections of climate, hydrology, ecosystems, and water resources management.

Appendix

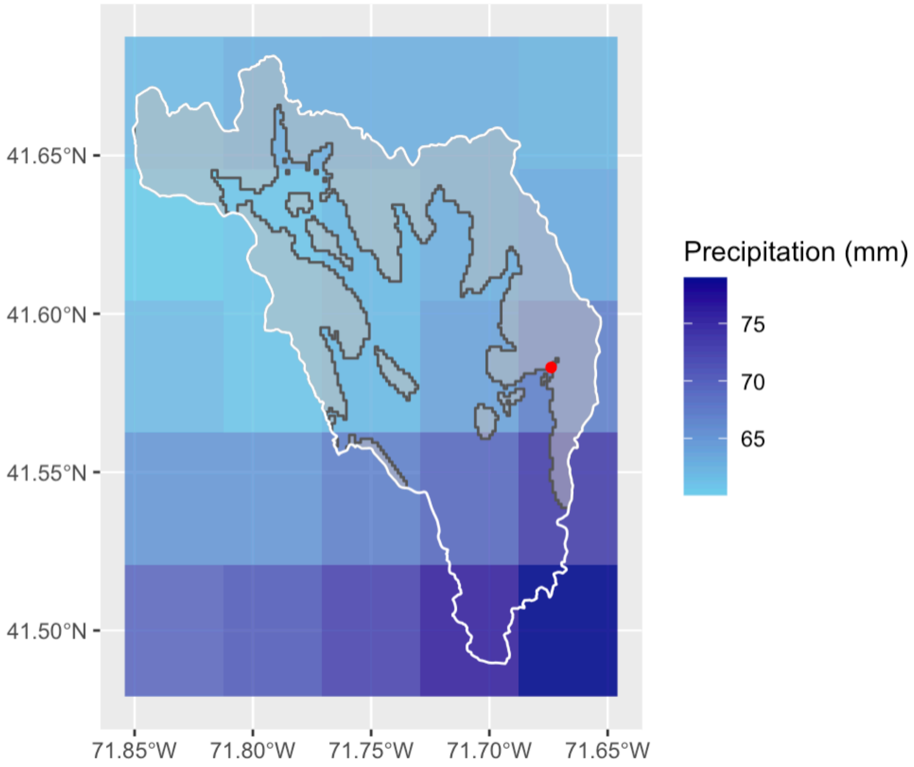

Figure 3-A.1 An example of the grid selection process for hydrologic predictors. The white polygon represents the HUC12 watershed, the red dot represents the lake, and the gray represents areas of the watershed with elevation exceeding the lake surface.

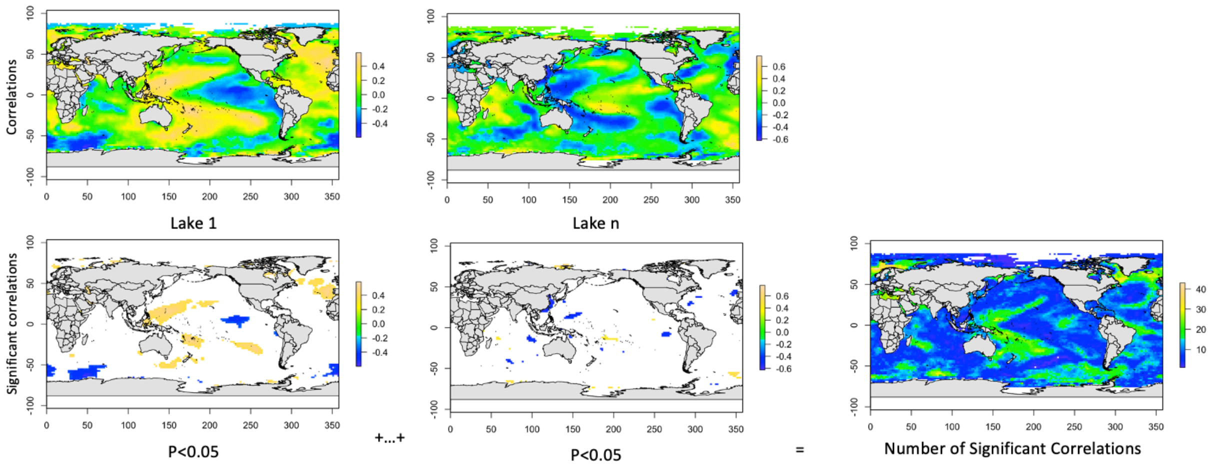

Figure 3-A.2 Sea surface temperature grid selection process; example from the Northeast region.

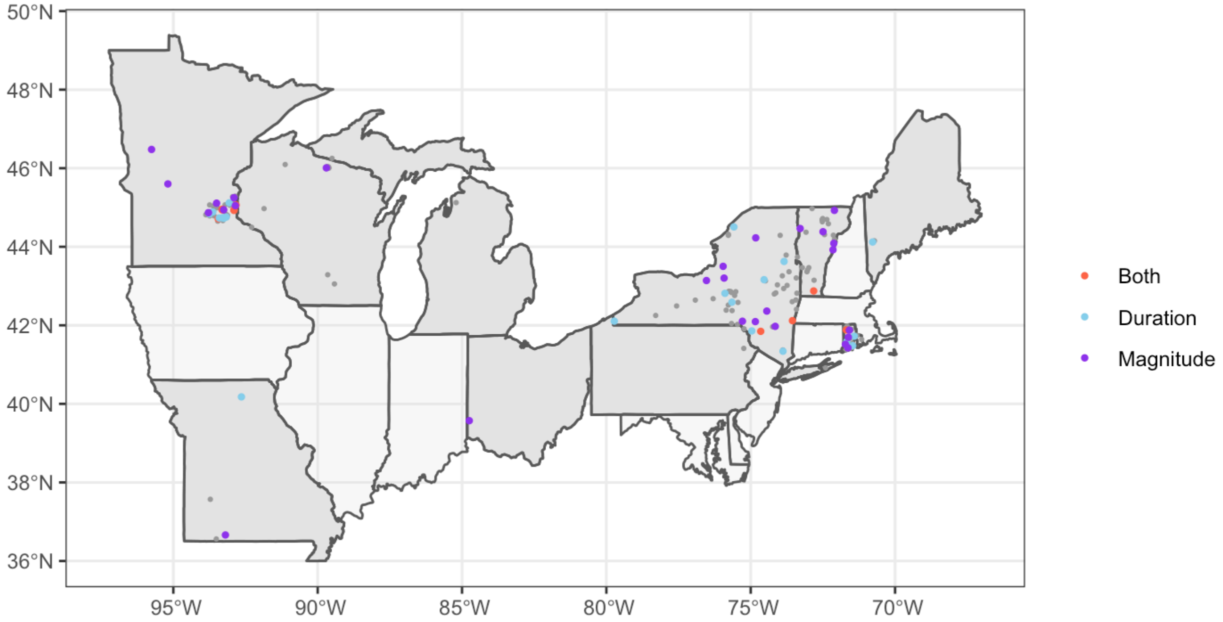

Figure 3-A.3 Lakes for which at least one hydroclimatic predictor is statistically significantly correlated with *magnitude*, *duration*, or both. Other lakes are shown in gray.

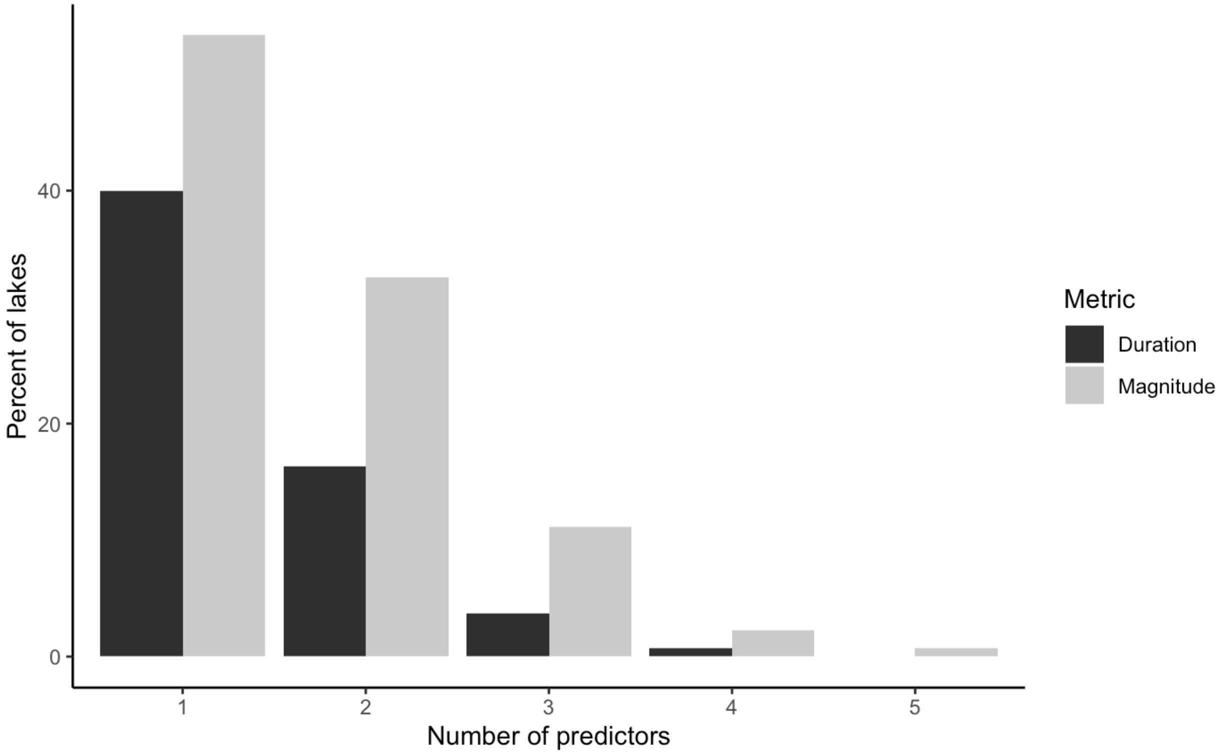

Figure 3-A.4 The number of predictors retained for each *magnitude* and *duration* lake model.

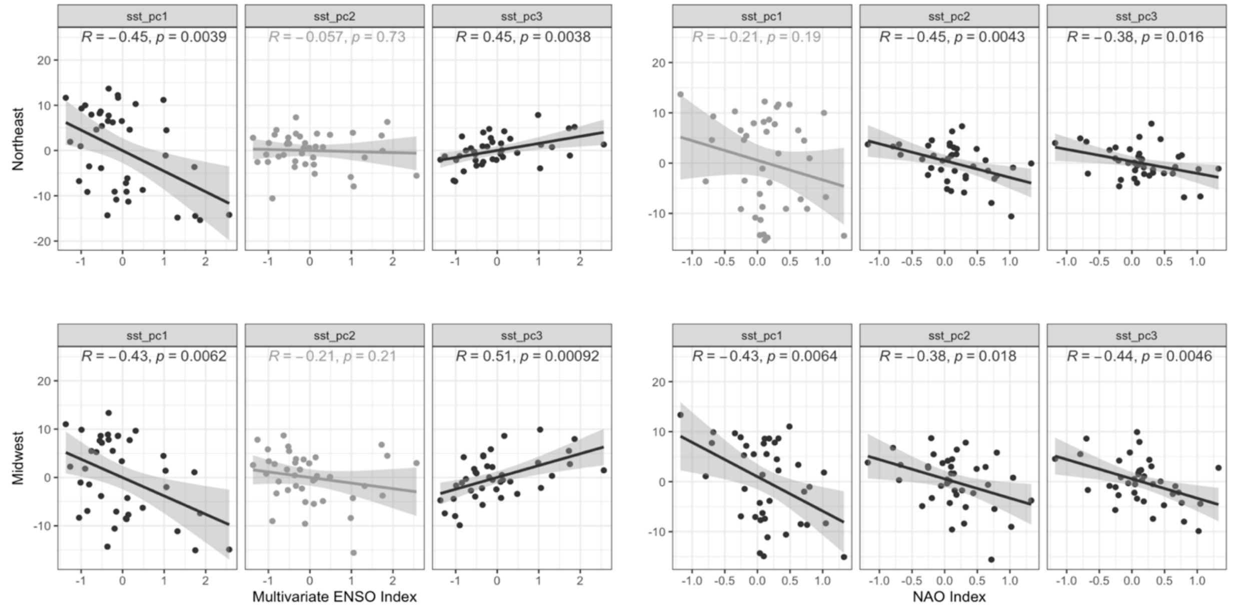

Figure 3-A.5 Correlation between selected SST principal components and NAO and MEI indices for the Northeast and Midwest U.S.

References

Alawadi, F. (2010). Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI). In *Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010* (Vol. 7825, pp. 45–58). SPIE.

Alexander, S., Wu, S., & Block, P. (2019). Model Selection Based on Sectoral Application Scale for Increased Value of Hydroclimate-Prediction Information. *Journal of Water Resources Planning and Management*, *145*(5), 04019006. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001044

Allan, R., & Ansell, T. (2006). A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *Journal of Climate*, *19*(22), 5816–5842.

Anneville, O., Domaizon, I., Kerimoglu, O., Rimet, F., & Jacquet, S. (2015). Blue-green algae in a "Greenhouse Century"? New insights from field data on climate change impacts on cyanobacteria abundance. *Ecosystems*, *18*(3), 441–458.

Aoki, I. (1989). Holological study of lakes from an entropy viewpoint-lake Mendota. *Ecological Modelling*, *45*(2), 81–93.

Arhonditsis, G. B., Winder, M., Brett, M. T., & Schindler, D. E. (2004). Patterns and mechanisms of phytoplankton variability in Lake Washington (USA). *Water Research*, *38*(18), 4013–4027.

Baharom, M., Ahmad, N., Hod, R., & Abdul Manaf, M. R. (2022). Dengue early warning system as outbreak prediction tool: A systematic review. *Risk Management and Healthcare Policy*, 871–886.

Baker, S. A., Wood, A. W., & Rajagopalan, B. (2019). Developing subseasonal to seasonal climate forecast products for hydrology and water management. *JAWRA Journal of the*

American Water Resources Association, *55*(4), 1024–1037.

Barnett, T. P. (1981). Statistical prediction of North American air temperatures from Pacific predictors. *Monthly Weather Review*, *109*(5), 1021–1041.

Barnston, A. G. (1994). Linear statistical short-term climate predictive skill in the Northern Hemisphere. *Journal of Climate*, *7*(10), 1513–1564. https://doi.org/10.1175/1520-0442(1994)007

Batur, E., & Maktav, D. (2018). Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(5), 2983–2989.

Beal, M. R. W., O'Reilly, B., Hietpas, K. R., & Block, P. (2021). Development of a sub-seasonal cyanobacteria prediction model by leveraging local and global scale predictors. *Harmful Algae*, *108*, 102100.

Beal, M. R. W., O'Reilly, B. E., Soley, C. K., Hietpas, K. R., & Block, P. J. (2022). Variability of summer cyanobacteria abundance: can season-ahead forecasts improve beach management? *Lake and Reservoir Management*, 1–16.

Beck, L. R., Rodriguez, M. H., Dister, S. W., Rodriguez, A. D., Rejmankova, E., Ulloa, A., et al. (1994). Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *The American Journal of Tropical Medicine and Hygiene*, *51*(3), 271–280.

Beck, R., Zhan, S., Liu, H., Tong, S., Yang, B., Xu, M., et al. (2016). Comparison of satellite reflectance algorithms for estimating chlorophyll-a in a temperate reservoir using coincident hyperspectral aircraft imagery and dense coincident surface observations. *Remote Sensing of Environment*, *178*, 15–30.

Beck, R., Xu, M., Zhan, S., Liu, H., Johansen, R. A., Tong, S., et al. (2017). Comparison of Satellite Reflectance Algorithms for Estimating Phycocyanin Values and Cyanobacterial Total Biovolume in a Temperate Reservoir Using Coincident Hyperspectral Aircraft Imagery and Dense Coincident Surface Observations. *Remote Sensing* . https://doi.org/10.3390/rs9060538

Becker, R. H., Sultan, M. I., Boyer, G. L., Twiss, M. R., & Konopko, E. (2009). Mapping cyanobacterial blooms in the Great Lakes using MODIS. *Journal of Great Lakes Research*, *35*(3), 447–453.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31.

Betz, C., & Genskow, K. (2012). Farm Practices in the Lake Mendota Watershed: A Comparative Analysis of 1996 and 2011. U of Wisconsin-Extension Environmental Resources Center. http://www ….

Beversdorf, L. J., Miller, T. R., & McMahon, K. D. (2013). The role of nitrogen fixation in cyanobacterial bloom toxicity in a temperate, eutrophic lake. *PloS One*, *8*(2), e56103.

Binding, C. E., Greenberg, T. A., & Bukata, R. P. (2012). An analysis of MODIS-derived algal and mineral turbidity in Lake Erie. *Journal of Great Lakes Research*, *38*(1), 107–116.

Block, P. J., Souza Filho, F. A., Sun, L., & Kwon, H. (2009). A streamflow forecasting framework using multiple climate and hydrological models 1. *JAWRA Journal of the American Water Resources Association*, *45*(4), 828–843.

Bortman, M. (1999). Elaboración de corredores o canales endémicos mediante planillas de cálculo. *Revista Panamericana de Salud Pública*, *5*, 1–8.

Boucher, J., Weathers, K. C., Norouzi, H., & Steele, B. (2018). Assessing the effectiveness of

Landsat 8 chlorophyll a retrieval algorithms for regional freshwater monitoring. *Ecological Applications*, *28*(4), 1044–1054.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brezonik, P. L., & Lee, G. F. (1968). Dentrification as a nitrogen sink in Lake Mendota, Wisconsin. *Environmental Science & Technology*, *2*(2), 120–125.

Brooks, B. W., Lazorchak, J. M., Howard, M. D. A., Johnson, M. V, Morton, S. L., Perkins, D. A. K., et al. (2016). Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environmental Toxicology and Chemistry*, *35*(1), 6–13.

Brunet, G., Shapiro, M., Hoskins, B., Moncrieff, M., Dole, R., Kiladis, G. N., et al. (2010). Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bulletin of the American Meteorological Society*, *91*(10), 1397–1406.

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

Cabrera, C. V. P., & Selvaraj, J. J. (2020). Geographic shifts in the bioclimatic suitability for Aedes aegypti under climate change scenarios in Colombia. *Heliyon*, *6*(1).

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., & Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. *Remote Sensing of Environment*, *248*, 111974.

Carey, C. C., Ibelings, B. W., Hoffmann, E. P., Hamilton, D. P., & Brookes, J. D. (2012). Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate. *Water Research*, *46*(5), 1394–1407.

Carlson, R. E. (1977). A trophic state index for lakes 1. *Limnology and Oceanography*, *22*(2), 361–369.

Carmichael, W. W. (1994). The toxins of cyanobacteria. *Scientific American*, *270*(1), 78–86.

Carmichael, W. W. (2001). Health effects of toxin-producing cyanobacteria:"The CyanoHABs." *Human and Ecological Risk Assessment: An International Journal*, *7*(5), 1393–1407.

Carmichael, W. W., & Boyer, G. L. (2016). Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes. *Harmful Algae*, *54*, 194–212. https://doi.org/https://doi.org/10.1016/j.hal.2016.02.002

Carpenter, S R, Kitchell, J. F., Hodgson, J. R., Cochran, P. A., Elser, J. J., Elser, M. M., et al. (1987). Regulation of lake primary productivity by food web structure. *Ecology*, *68*(6), 1863–1876.

Carpenter, Stephen R, Lathrop, R. C., Nowak, P., Bennett, E. M., Reed, T., & Soranno, P. A. (2006). The ongoing experiment: restoration of Lake Mendota and its watershed. *Long-Term Dynamics of Lakes in the Landscape. Oxford University Press, New York*, 236–256.

Carpenter, Stephen R, Booth, E. G., Kucharik, C. J., & Lathrop, R. C. (2015). Extreme daily loads: role in annual phosphorus input to a north temperate lake. *Aquatic Sciences*, *77*(1), 71–79.

Carpenter, Stephen R, Booth, E. G., & Kucharik, C. J. (2018). Extreme precipitation and phosphorus loads from two agricultural watersheds. *Limnology and Oceanography*, *63*(3), 1221–1233.

Castillo, M. M., Allan, J. D., & Brunzell, S. (2000). *Nutrient concentrations and discharges in a Midwestern agricultural catchment*. Wiley Online Library.

Center, M. R. C. (2016). Midwest Climate: El Niño.

Chebud, Y., Naja, G. M., Rivero, R. G., & Melesse, A. M. (2012). Water quality monitoring using remote sensing and an artificial neural network. *Water, Air, & Soil Pollution*, *223*(8), 4875–4887.

Chen, Q., Guan, T., Yun, L., Li, R., & Recknagel, F. (2015). Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials. *Harmful Algae*, *43*, 58–65.

Chiew, F. H. S., Zhou, S. L., & McMahon, T. A. (2003). Use of seasonal streamflow forecasts in water resources management. *Journal of Hydrology*, *270*(1–2), 135–144.

Cho, J., Shin, C.-M., Choi, H.-K., Kim, K.-H., & Choi, J.-Y. (2016). Development of an integrated method for long-term water quality prediction using seasonal climate forecast. *Proceedings of the International Association of Hydrological Sciences*, *374*, 175–185.

Claypool, A. L., Brandeau, M. L., & Goldhaber-Fiebert, J. D. (2021). Prevention and control of dengue and chikungunya in Colombia: A cost-effectiveness analysis. *PLoS Neglected Tropical Diseases*, *15*(12), e0010086.

Cottingham, K. L., Rusak, J. A., & Leavitt, P. R. (2000). Increased ecosystem variability and reduced predictability following fertilisation: Evidence from palaeolimnology. *Ecology Letters*, *3*(4), 340–348.

Craig, M. H., Snow, R. W., & le Sueur, D. (1999). A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitology Today*, *15*(3), 105–111.

Dalton, C. (2021, October 27). Kaptur Announces $1.77 Million for Great Lakes Harmful Algal Bloom Research Projects. *House Great Lakes Task Force*. Retrieved from https://s3.documentcloud.org/documents/21094180/press-release-kaptur-announces-17m-for-great-lakes-harmful-algal-bloom-research-projects.pdf

Dee, D. P., & Da Silva, A. M. (1998). Data assimilation in the presence of forecast bias. *Quarterly Journal of the Royal Meteorological Society*, *124*(545), 269–295.

Dekker, A. G. (1993). Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing.

Delorit, J., Gonzalez Ortuya, E. C., & Block, P. (2017). Evaluation of model-based seasonal streamflow and water allocation forecasts for the Elqui Valley, Chile. *Hydrology and Earth System Sciences*, *21*(9), 4711–4725.

Deng, Jiancai, Chen, F., Liu, X., Peng, J., & Hu, W. (2016). Horizontal migration of algal patches associated with cyanobacterial blooms in an eutrophic shallow lake. *Ecological Engineering*, *87*, 185–193.

Deng, Jianming, Qin, B., Paerl, H. W., Zhang, Y., Ma, J., & Chen, Y. (2014). Earlier and warmer springs increase cyanobacterial (Microcystis spp.) blooms in subtropical Lake Taihu, China. *Freshwater Biology*, *59*(5), 1076–1085.

Dillon, P. J., & Rigler, F. H. (1974). The phosphorus-chlorophyll relationship in lakes 1, 2. *Limnology and Oceanography*, *19*(5), 767–773.

Dodds, W. K., Bouska, W. W., Eitzmann, J. L., Pilger, T. J., Pitts, K. L., Riley, A. J., et al. (2009). Eutrophication of US freshwaters: analysis of potential economic damages. ACS Publications.

Downing, J. A., Watson, S. B., & McCauley, E. (2001). Predicting cyanobacteria dominance in lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(10), 1905–1908.

Duarte, J. L., Diaz-Quijano, F. A., Batista, A. C., & Giatti, L. L. (2019). Climatic variables associated with dengue incidence in a city of the Western Brazilian Amazon region. *Revista*

Da Sociedade Brasileira de Medicina Tropical, *52*.

Edmondson, W. T., & Lehman, J. T. (1981). The effect of changes in the nutrient income on the condition of Lake Washington 1. *Limnology and Oceanography*, *26*(1), 1–29.

Elliott, J. A., Jones, I. D., & Thackeray, S. J. (2006). Testing the sensitivity of phytoplankton communities to changes in water temperature and nutrient load, in a temperate lake. *Hydrobiologia*, *559*(1). https://doi.org/10.1007/s10750-005-1233-y

Elliott, J Alex. (2012). Is the future blue-green? A review of the current model predictions of how climate change could affect pelagic freshwater cyanobacteria. *Water Research*, *46*(5), 1364–1371.

Enfield, D. B., Mestas-Nuñez, A. M., & Trimble, P. J. (2001). The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US. *Geophysical Research Letters*, *28*(10), 2077–2080.

Epp, G. T. (1996). Grazing on filamentous cyanobacteria by Daphnia pulicaria. *Limnology and Oceanography*, *41*(3), 560–567.

Eppley, R. W. (1972). Temperature and phytoplankton growth in the sea. *Fish. Bull*, *70*(4), 1063–1085.

Epstein, E. S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology and Climatology*, *8*(6), 985–987. https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2

Farquhar, G. J. (2010). Leachate: production and characterization. *Can. J. Civ. Eng.*, *16*(3), 317–325. https://doi.org/10.1139/l89-057

Fee, E. J., Hecky, R. E., Regehr, G. W., Hendzel, L. L., & Wilkinson, P. (1994). Effects of lake size on nutrient availability in the mixed layer during summer stratification. *Canadian Journal of Fisheries and Aquatic Sciences*, *51*(12), 2756–2768.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, *2*(1), 150066. https://doi.org/10.1038/sdata.2015.66

Gallina, N., Anneville, O., & Beniston, M. (2011). Impacts of extreme air temperatures on cyanobacteria in five deep peri-Alpine lakes. *Journal of Limnology*, *70*(2), 186.

Garnache, C., Swinton, S. M., Herriges, J. A., Lupi, F., & Stevenson, R. J. (2016). Solving the phosphorus pollution puzzle: Synthesis and directions for future research. *American Journal of Agricultural Economics*, *98*(5), 1334–1359.

Gholizadeh, M. H., Melesse, A. M., & Reddi, L. (2016). A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors*, *16*(8), 1298.

Giannini, A., Kushnir, Y., & Cane, M. A. (2000). Interannual variability of Caribbean rainfall, ENSO, and the Atlantic Ocean. *Journal of Climate*, *13*(2), 297–311.

Gilerson, A. A., Gitelson, A. A., Zhou, J., Gurlin, D., Moses, W., Ioannou, I., & Ahmed, S. A. (2010). Algorithms for remote estimation of chlorophyll-a in coastal and inland waters using red and near infrared bands. *Optics Express*, *18*(23), 24109–24125.

Gitelson, A. A., Gritz, Y., & Merzlyak, M. N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, *160*(3), 271–282.

Giuliani, M., Zaniolo, M., Castelletti, A., Davoli, G., & Block, P. (2019). Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. *Water Resources Research*, *55*(11), 9133–9147.

Giuliani, M. Zaniolo, M. Castelletti, A., Block, P., Zimmerman B., Carlino, A. Amaranto, A.

Climate State Intelligence. 2019b. https://github.com/mxgiuliani00/CSI

Glavan, M., Ceglar, A., & Pintar, M. (2015). Assessing the impacts of climate change on water quantity and quality modelling in small Slovenian Mediterranean catchment–lesson for policy and decision makers. *Hydrological Processes*, *29*(14), 3124–3144.

Glibert, P. M., & Burkholder, J. M. (2006). The complex relationships between increases in fertilization of the earth, coastal eutrophication and proliferation of harmful algal blooms. In *Ecology of harmful algae* (pp. 341–354). Springer.

Gobler, C. J., Burkholder, J. M., Davis, T. W., Harke, M. J., Johengen, T., Stow, C. A., & Van de Waal, D. B. (2016). The dual role of nitrogen supply in controlling the growth and toxicity of cyanobacterial blooms. *Harmful Algae*, *54*, 87–97.

Gower, J., King, S., Borstad, G., & Brown, L. (2005). Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer. *International Journal of Remote Sensing*, *26*(9), 2005–2012.

Gutierrez-Barbosa, H., Medina-Moreno, S., Zapata, J. C., & Chua, J. V. (2020). Dengue infections in Colombia: epidemiological trends of a hyperendemic country. *Tropical Medicine and Infectious Disease*, *5*(4), 156.

Guzman, M. G., Halstead, S. B., Artsob, H., Buchy, P., Farrar, J., Gubler, D. J., et al. (2010). Dengue: a continuing global threat. *Nature Reviews Microbiology*, *8*(Suppl 12), S7–S16.

Håkanson, L., Bryhn, A. C., & Hytteborn, J. K. (2007). On the issue of limiting nutrient and predictions of cyanobacteria in aquatic systems. *Science of the Total Environment*, *379*(1), 89–108.

Hamilton, D. P., Salmaso, N., & Paerl, H. W. (2016). Mitigating harmful cyanobacterial blooms: strategies for control of nitrogen and phosphorus loads. *Aquatic Ecology*, *50*, 351–366.

Han, L., & Jordan, K. J. (2005). Estimating and mapping chlorophyll-a concentration in Pensacola Bay, Florida using Landsat ETM+ data. *International Journal of Remote Sensing*, *26*(23), 5245–5254.

Hansen, J. W., Mason, S. J., Sun, L., & Tall, A. (2011). Review of seasonal climate forecasting for agriculture in sub-Saharan Africa. *Experimental Agriculture*, *47*(2), 205–240.

Haygarth, P. M., & Jarvis, S. C. (1997). Soil derived phosphorus in surface runoff from grazed grassland lysimeters. *Water Research*, *31*(1), 140–148. https://doi.org/https://doi.org/10.1016/S0043-1354(99)80002-5

Heidke, P. (1926). Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geogr. Ann.*, *8*, 301–349.

Helsel, D. R., & Hirsch, R. M. (1992). *Statistical methods in water resources* (Vol. 49). Elsevier.

Hill, P. R., Kumar, A., Temimi, M., & Bull, D. R. (2020). HABNet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3229–3239.

Ho, J. C., Michalak, A. M., & Pahlevan, N. (2019). Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*, *574*(7780), 667–670.

Houlahan, J. E., McKinney, S. T., Anderson, T. M., & McGill, B. J. (2017). The priority of prediction in ecological understanding. *Oikos*, *126*(1), 1–7.

Hu, C. (2009). A novel ocean color index to detect floating algae in the global oceans. *Remote Sensing of Environment*, *113*(10), 2118–2129.

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. H., & Visser, P. M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology*, *16*(8), 471–483.

Hunter, P. R. (2003). Climate change and waterborne and vector-borne disease. *Journal of*

Applied Microbiology, *94*(s1), 37–46.

Impacts, W. I. on C. C. (2011). Wisconsin's changing climate: impacts and adaptation. Nelson Institute for Environmental Studies, University of Wisconsin–Madison ….

International Joint Commission. (2014). *A Balanced Diet for Lake Erie*.

IPCC, F. C. B. (2012). Managing the risks of extreme events and disasters to advance climate change adaptation. *A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*, *582*.

Joehnk, K. D., Huisman, J. E. F., Sharples, J., Sommeijer, B. E. N., Visser, P. M., & Stroom, J. M. (2008). Summer heatwaves promote blooms of harmful cyanobacteria. *Global Change Biology*, *14*(3), 495–512.

Johansson, M. A., Dominici, F., & Glass, G. E. (2009). Local and global effects of climate on dengue transmission in Puerto Rico. *PLoS Neglected Tropical Diseases*, *3*(2), e382.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990–993.

Justić, D., Rabalais, N. N., & Turner, R. E. (2005). Coupling between climate variability and coastal eutrophication: evidence and outlook for the northern Gulf of Mexico. *Journal of Sea Research*, *54*(1), 25–35.

Kahya, E., & Dracup, J. A. (1993). US streamflow patterns in relation to the El Niño/Southern Oscillation. *Water Resources Research*, *29*(8), 2491–2503.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Kane, D. D., Conroy, J. D., Peter Richards, R., Baker, D. B., & Culver, D. A. (2014). Re-eutrophication of Lake Erie: Correlations between tributary nutrient loads and phytoplankton biomass. *Journal of Great Lakes Research*, *40*(3), 496–501. https://doi.org/https://doi.org/10.1016/j.jglr.2014.04.004

Kasprzak, P. H., & Lathrop, R. C. (1997). Influence of two Daphnia species on summer phytoplankton assemblages from eutrophic lakes. *Journal of Plankton Research*, *19*(8), 1025–1044.

Katzelnick, L. C., Bos, S., & Harris, E. (2020). Protective and enhancing interactions among dengue viruses 1-4 and Zika virus. *Current Opinion in Virology*, *43*, 59–70.

Kavanaugh, K. E., Derner, K., Fisher, K. M., David, E. E. P., Urizar, C., & Merlini, R. (2013). Assessment of the eastern Gulf of Mexico Harmful Algal Bloom Operational Forecast System (GOMX HAB-OFS): A comparative analysis of forecast skill and utilization from October 1, 2004 to April 30, 2008.

Kim, Y. H., Son, S., Kim, H.-C., Kim, B., Park, Y.-G., Nam, J., & Ryu, J. (2020). Application of satellite remote sensing in monitoring dissolved oxygen variabilities: A case study for coastal waters in Korea. *Environment International*, *134*, 105301.

Kleinman, P. J. A., Srinivasan, M. S., Dell, C. J., Schmidt, J. P., Sharpley, A. N., & Bryant, R. B. (2006). Role of rainfall intensity and hydrology in nutrient transport via surface runoff.

Konopka, A., & Brock, T. D. (1978). Effect of temperature on blue-green algae (cyanobacteria) in Lake Mendota. *Applied and Environmental Microbiology*, *36*(4), 572–576.

Krishnamurthy, L., Vecchi, G. A., Msadek, R., Wittenberg, A., Delworth, T. L., & Zeng, F. (2015). The seasonality of the Great Plains low-level jet and ENSO relationship. *Journal of Climate*, *28*(11), 4525–4544.

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informaticae*, *101*(4), 271–285.

Lala, J., Tilahun, S., & Block, P. (2020). Predicting rainy season onset in the Ethiopian Highlands for agricultural planning. *Journal of Hydrometeorology*, *21*(7), 1675–1688.

Lampert, W., Fleckner, W., Rai, H., & Taylor, B. E. (1986). Phytoplankton control by grazing zooplankton: A study on the spring clear-water phase 1. *Limnology and Oceanography*, *31*(3), 478–490.

Lathrop, R C, Carpenter, S. R., Stow, C. A., Soranno, P. A., & Panuska, J. C. (1998). Phosphorus loading reductions needed to control blue-green algal blooms in Lake Mendota. *Canadian Journal of Fisheries and Aquatic Sciences*, *55*(5), 1169–1178. https://doi.org/10.1139/f97-317

Lathrop, Richard C. (2007). Perspectives on the eutrophication of the Yahara lakes. *Lake and Reservoir Management*, *23*(4), 345–365.

Lathrop, Richard C, & Carpenter, S. R. (1992). Phytoplankton and their relationship to nutrients. In *Food Web Management* (pp. 97–126). Springer.

Lathrop, Richard C, & Carpenter, S. R. (2014). Water quality implications from three decades of phosphorus loads and trophic dynamics in the Yahara chain of lakes. *Inland Waters*, *4*(1), 1–14.

Lathrop, Richard C, Carpenter, S. R., & Robertson, D. M. (1999). Summer water clarity responses to phosphorus, Daphnia grazing, and internal mixing in Lake Mendota. *Limnology and Oceanography*, *44*(1), 137–146.

Lee, D., Ward, P. J., & Block, P. (2018). Identification of symmetric and asymmetric responses in seasonal streamflow globally to ENSO phase. *Environmental Research Letters*, *13*(4), 44031.

Legler, D. M., Bryant, K. J., & O'brien, J. J. (1999). Impact of ENSO-related climate anomalies on crop yields in the US. *Climatic Change*, *42*, 351–375.

Lenharo, M. (2023). Dengue is breaking records in the Americas-what's behind the surge? *Nature*.

Lenters, J. D., & Cook, K. H. (1997). On the origin of the Bolivian high and related circulation features of the South American climate. *Journal of the Atmospheric Sciences*, *54*(5), 656–678.

León-Muñoz, J., Urbina, M. A., Garreaud, R., & Iriarte, J. L. (2018). Hydroclimatic conditions trigger record harmful algal bloom in western Patagonia (summer 2016). *Scientific Reports*, *8*(1), 1–10.

Liu, R., Wang, J., Shi, J., Chen, Y., Sun, C., Zhang, P., & Shen, Z. (2014). Runoff characteristics and nutrient loss mechanism from plain farmland under simulated rainfall conditions. *Science of the Total Environment*, *468*, 1069–1077.

Liu, X., Feng, J., & Wang, Y. (2019). Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales. *Science of the Total Environment*, *648*, 472–480.

Lowe, R., Lee, S. A., O'Reilly, K. M., Brady, O. J., Bastos, L., Carrasco-Escobar, G., et al. (2021). Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. *The Lancet Planetary Health*, *5*(4), e209–e219.

Magee, M. R., McIntyre, P. B., Hanson, P. C., & Wu, C. H. (2019). Drivers and management implications of long-term Cisco oxythermal habitat decline in Lake Mendota, WI. *Environmental Management*, *63*(3), 396–407.

Magnuson, J. J., Carpenter, S. R., & Stanley, E. H. (2019). North Temperate Lakes LTER: Zooplankton-Madison Lakes Area 1997-current. Retrieved from https://doi.org/10.6073/pasta/8b265c0300252c87805f26f41e174aa4.

Magnuson, J. J., Carpenter, S. R., & Stanley, E. H. (2022). North Temperate Lakes LTER: Phytoplankton-Madison Lakes Area 1995-current.

Magnuson, J. J., Carpenter, S. R., & Stanley, E. H. (2023a). North Temperate Lakes LTER: Chemical Limnology of Primary Study Lakes: Nutrients, pH and Carbon 1981-current.

Magnuson, J. J., Carpenter, S. R., & Stanley, E. H. (2023b). North Temperate Lakes LTER: High Frequency Water Temperature Data-Lake Mendota Buoy 2006-current.

Mander, Ü., Kull, A., Kuusemets, V., & Tamm, T. (2000). Nutrient runoff dynamics in a rural catchment: Influence of land-use changes, climatic fluctuations and ecotechnological measures. *Ecological Engineering*, *14*(4), 405–417. https://doi.org/https://doi.org/10.1016/S0925-8574(99)00064-6

Markowski, G. R., & North, G. R. (2003). Climatic influence of sea surface temperature: Evidence of substantial precipitation correlation and predictability. *Journal of Hydrometeorology*, *4*(5), 856–877. https://doi.org/10.1175/1525-7541(2003)004<0856:CIOSST>2.0.CO;2

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817.

Me, W., Hamilton, D. P., McBride, C. G., Abell, J. M., & Hicks, B. J. (2018). Modelling hydrology and water quality in a mixed land use catchment and eutrophic lake: Effects of nutrient load reductions and climate change. *Environmental Modelling & Software*, *109*, 114–133.

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., et al. (2012). Global historical climatology network-daily (GHCN-Daily), Version 3. *NOAA National Climatic Data Center*, *10*(10.7289), V5D21VHZ.

Michalak, A. M. (2016). Study role of climate change in extreme threats to water quality. *Nature*, *535*(7612), 349–350.

Midekisa, A., Senay, G., Henebry, G. M., Semuniguse, P., & Wimberly, M. C. (2012). Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. *Malaria Journal*, *11*(1), 1–10.

Mishra, S., & Mishra, D. R. (2012). Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sensing of Environment*, *117*, 394–406. https://doi.org/https://doi.org/10.1016/j.rse.2011.10.016

Mocko, D., NASA/GSFC/HSL. 2013. NLDAS Noah Land Surface Model L4 Monthly Climatology 0.125 x 0.125 degree V002, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/U5BAYF8R76IK. Accessed: 2020-09-16.

Morse, N. B., & Wollheim, W. M. (2014). Climate variability masks the impacts of land use change on nutrient export in a suburbanizing watershed. *Biogeochemistry*, *121*, 45–59.

Moss, B. (2012). Cogs in the endless machine: lakes, climate change and nutrient cycles: a review. *Science of the Total Environment*, *434*, 130–142.

Motew, M., Chen, X., Booth, E. G., Carpenter, S. R., Pinkas, P., Zipper, S. C., et al. (2017). The influence of legacy P on lake water quality in a midwestern agricultural watershed. *Ecosystems*, *20*, 1468–1482.

Muñoz, Á. G., Chourio, X., Rivière-Cinnamond, A., Diuk-Wasser, M. A., Kache, P. A., Mordecai, E. A., et al. (2020). Ae DES: a next-generation monitoring and forecasting system for environmental suitability of Aedes-borne disease transmission. *Scientific Reports*, *10*(1), 12640.

NCEI. (2022). US billion-dollar weather and climate disasters. NCEI Washington DC, USA.

O'Neil, J. M., Davis, T. W., Burford, M. A., & Gobler, C. J. (2012). The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. *Harmful Algae*, *14*, 313–334.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., et al. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research: Oceans*, *103*(C11), 24937–24953.

ORNL DAAC. 2020. MODIS and VIIRS Land Products Global Subsetting and Visualization Tool. ORNL DAAC, Oak Ridge, Tennessee, USA. Accessed September 16, 2020. Subset obtained for MCD43A4.006 MODIS Nadir BRDF-Adjusted Reflectance Daily 500m product at 43.0989,-89.4055, time period: 6-1-2020 to 6-26-2020, and subset size: 0.5 x 0.5km. https://doi.org/10.3334/ORNLDAAC/1379. Accessed 2020-10-02.

Obenour, D. R., Gronewold, A. D., Stow, C. A., & Scavia, D. (2014). Using a B ayesian hierarchical model to improve L ake E rie cyanobacteria bloom forecasts. *Water Resources Research*, *50*(10), 7847–7860.

Ocampo, C. B., Mina, N. J., Carabalí, M., Alexander, N., & Osorio, L. (2014). Reduction in dengue cases observed during mass control of Aedes (Stegomyia) in street catch basins in an endemic urban area in Colombia. *Acta Tropica*, *132*, 15–22.

Ostfeld, A., Tubaltzev, A., Rom, M., Kronaveter, L., Zohary, T., & 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), 4014069.

Oyama, Y., Fukushima, T., Matsushita, B., Matsuzaki, H., Kamiya, K., & Kobinata, H. (2015). Monitoring levels of cyanobacterial blooms using the visual cyanobacteria index (VCI) and floating algae index (FAI). *International Journal of Applied Earth Observation and Geoinformation*, *38*, 335–348.

Paerl, H. W. (1988). Nuisance phytoplankton blooms in coastal, estuarine, and inland waters 1. *Limnology and Oceanography*, *33*(4part2), 823–843.

Paerl, H. W. (2017). Controlling cyanobacterial harmful blooms in freshwater ecosystems. *Microbial Biotechnology*, *10*(5), 1106–1110.

Paerl, H. W., & Huisman, J. (2008). Blooms like it hot. *Science*, *320*(5872), 57–58.

Paerl, H. W., & Otten, T. G. (2016). Duelling 'CyanoHABs': unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-N2-fixing harmful cyanobacteria. *Environmental Microbiology*, *18*(2), 316–324.

Paerl, H. W., & Paul, V. J. (2012). Climate change: links to global expansion of harmful cyanobacteria. *Water Research*, *46*(5), 1349–1363.

Paerl, H. W., Fulton, R. S., Moisander, P. H., & Dyble, J. (2001). Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *TheScientificWorldJournal*, *1*, 76–113.

Pahlevan, N., Chittimalli, S. K., Balasubramanian, S. V, & Vellucci, V. (2019). Sentinel-2/Landsat-8 product consistency and implications for monitoring aquatic systems. *Remote Sensing of Environment*, *220*, 19–29.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., et al. (2020). Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sensing of Environment*, *240*, 111604. https://doi.org/https://doi.org/10.1016/j.rse.2019.111604

Park, J.-Y., Stock, C. A., Dunne, J. P., Yang, X., & Rosati, A. (2019). Seasonal to multiannual marine ecosystem prediction with a global Earth system model. *Science*, *365*(6450), 284–288.

Patel, J., & Parshina-Kottas, Y. (2017, October 3). Miles of Algae Covering Lake Erie. *New York Times*. Retrieved from https://www.nytimes.com/interactive/2017/10/03/science/earth/lake-erie.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez-González, R., Sòria-Perpinyà, X., Soria, J. M., Delegido, J., Urrego, P., Sendra, M. D., et al. (2021). Phycocyanin Monitoring in Some Spanish Water Bodies with Sentinel-2 Imagery. *Water*, *13*(20), 2866.

Persaud, A. D., Paterson, A. M., Dillon, P. J., Winter, J. G., Palmer, M., & Somers, K. M. (2015). Forecasting cyanobacteria dominance in Canadian temperate lakes. *Journal of Environmental Management*, *151*, 343–352.

Poveda, Germain, Graham, N. E., Epstein, P. R., Rojas, W., Quiñones, M. L., Velez, I. D., & Martens, W. J. M. (2000). Climate and ENSO variability associated with vector-borne diseases in Colombia. *El Niño and the Southern Oscillation, Multiscale Variability and Global and Regional Impacts*, *1*, 183–204.

Poveda, Germán, Alvarez, D. M., & Rueda, O. A. (2011). Hydro-climatic variability over the Andes of Colombia associated with ENSO: a review of climatic processes and their impact on one of the Earth's most important biodiversity hotspots. *Climate Dynamics*, *36*, 2233–2249.

Qi, L., Hu, C., Duan, H., Cannizzaro, J., & Ma, R. (2014). A novel MERIS algorithm to derive cyanobacterial phycocyanin pigment concentrations in a eutrophic lake: Theoretical basis and practical considerations. *Remote Sensing of Environment*, *154*, 298–317.

Qian, S. S., Stow, C. A., Rowland, F. E., Liu, Q., Rowe, M. D., Anderson, E. J., et al. (2021). Chlorophyll a as an indicator of microcystin: Short-term forecasting and risk assessment in Lake Erie. *Ecological Indicators*, *130*, 108055.

Qin, B., Gao, G., Zhu, G., Zhang, Y., Song, Y., Tang, X., et al. (2013). Lake eutrophication and its ecosystem response. *Chinese Science Bulletin*, *58*(9), 961–970.

Reichwaldt, E. S., & Ghadouani, A. (2012). Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: between simplistic scenarios and complex dynamics. *Water Research*, *46*(5), 1372–1393.

Reutter, J. M., Ciborowski, J., DePinto, J., Bade, D., Baker, D., Bridgeman, T. B., et al. (2011). Lake Erie nutrient loading and harmful algal blooms: research findings and management implications. *Final Report of the Lake Erie Millennium Network Synthesis Team*.

Reynolds, C. S. (1984). *The ecology of freshwater phytoplankton*. Cambridge university press.

Rico-Mendoza, A., Alexandra, P.-R., Chang, A., Encinales, L., & Lynch, R. (2019). Co-circulation of dengue, chikungunya, and Zika viruses in Colombia from 2008 to 2018. *Revista Panamericana de Salud Pública*, *43*.

Rigosi, A., Carey, C. C., Ibelings, B. W., & Brookes, J. D. (2014). The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. *Limnology and Oceanography*, *59*(1), 99–114.

Rigosi, A., Hanson, P., Hamilton, D. P., Hipsey, M., Rusak, J. A., Bois, J., et al. (2015).

Determining the probability of cyanobacterial blooms: the application of Bayesian networks in multiple lake systems. *Ecological Applications*, *25*(1), 186–199.

Robarts, R. D., & Zohary, T. (1987). Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. *New Zealand Journal of Marine and Freshwater Research*, *21*(3), 391–399.

Robertson, D. (2016). Lake Mendota water temperature secchi depth snow depth ice thickness and meterological conditions 1894-2007.

Roelke, D., & Buyukates, Y. (2001). The diversity of harmful algal bloom-triggering mechanisms and the complexity of bloom initiation. *Human and Ecological Risk Assessment: An International Journal*, *7*(5), 1347–1362.

Ropelewski, C. F., & Halpert, M. S. (1986). North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Monthly Weather Review*, *114*(12), 2352–2362.

Ropelewski, C. F., & Halpert, M. S. (1987). Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, *115*(8), 1606–1626.

da Rosa Wieliczko, A., Crossetti, L. O., Cavalcanti, J. R., Hessel, M. S., da Motta-Marques, D., & Rodrigues, L. R. (2021). Meteorological drivers and ENSO influence on phytoplankton biomass dynamics in a shallow subtropical lake. *Environmental Monitoring and Assessment*, *193*(8), 536. https://doi.org/10.1007/s10661-021-09288-4

Rousseaux, C. S., Gregg, W. W., & Ott, L. (2021). Assessing the Skills of a Seasonal Forecast of Chlorophyll in the Global Pelagic Oceans. *Remote Sensing*, *13*(6), 1051.

Rousso, B. Z., Bertone, E., Stewart, R., & Hamilton, D. P. (2020). A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Research*, *182*, 115959.

Royer, T. V., David, M. B., & Gentry, L. E. (2006). Timing of riverine export of nitrate and phosphorus from agricultural watersheds in Illinois: Implications for reducing nutrient loading to the Mississippi River. *Environmental Science and Technology*, *40*(13), 4126–4131. https://doi.org/10.1021/es052573n

Rusak, J. A., Tanentzap, A. J., Klug, J. L., Rose, K. C., Hendricks, S. P., Jennings, E., et al. (2018). Wind and trophic status explain within and among-lake variability of algal biomass. *Limnology and Oceanography Letters*, *3*(6), 409–418.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., et al. (2020). Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews*, *205*, 103187.

Salas, E. A. L., Kumaran, S. S., Partee, E. B., Willis, L. P., & Mitchell, K. (2022). Potential of mapping dissolved oxygen in the Little Miami River using Sentinel-2 images and machine learning algorithms. *Remote Sensing Applications: Society and Environment*, *26*, 100759.

Salas, H. D., Poveda, G., Mesa, Ó. J., & Marwan, N. (2020). Generalized synchronization between ENSO and hydrological variables in Colombia: A recurrence quantification approach. *Frontiers in Applied Mathematics and Statistics*, *6*, 3.

Salmaso, N., Buzzi, F., Garibaldi, L., Morabito, G., & Simona, M. (2012). Effects of nutrient availability and temperature on phytoplankton development: a case study from large lakes south of the Alps. *Aquatic Sciences*, *74*(3), 555–570.

Sarachik, E. S., & Cane, M. A. (2010). *The El Nino-southern oscillation phenomenon.*

Cambridge University Press.

Sarnelle, O. (2007). Initial conditions mediate the interaction between Daphnia and bloom-forming cyanobacteria. *Limnology and Oceanography*, *52*(5), 2120–2127.

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, *5*(4), 570–575.

Schindler, D W. (1971). CARBON, NITROGEN, AND PHOSPHORUS AND THE EUTROPHICATION OF FRESHWATER LAKES 1. *Journal of Phycology*, *7*(4), 321–329.

Schindler, D W. (1978). Factors regulating phytoplankton production and standing crop in the world's freshwaters. *Limnology and Oceanography*, *23*(3), 478–486.

Schindler, David W. (1977). Evolution of phosphorus limitation in lakes: natural mechanisms compensate for deficiencies of nitrogen and carbon in eutrophied lakes. *Science*, *195*(4275), 260–262.

Schirrmeister, B. E., de Vos, J. M., Antonelli, A., & Bagheri, H. C. (2013). Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences*, *110*(5), 1791–1796.

Schueler, T. R. (1987). *Controlling urban runoff: A practical manual for planning and designing urban BMPs*. Metropolitan Washington Council of Governments Washington, DC.

Scordo, F., Lottig, N. R., Fiorenza, J. E., Culpepper, J., Simmons, J., Seitz, C., et al. (2022). Hydroclimate variability affects habitat-specific (open water and littoral) lake metabolism. *Water Resources Research*, e2021WR031094.

Sena, L., Deressa, W., & Ali, A. (2015). Correlation of climate variability and malaria: a retrospective comparative study, Southwest Ethiopia. *Ethiopian Journal of Health Sciences*, *25*(2), 129–138.

Sepulveda, L. S., & Vasilieva, O. (2016). Optimal control approach to dengue reduction and prevention in Cali, Colombia. *Mathematical Methods in the Applied Sciences*, *39*(18), 5475–5496.

Shabbar, A., & Skinner, W. (2004). Summer drought patterns in Canada and the relationship Toglobal sea surface temperatures. *Journal of Climate*, *17*(14), 2866–2880.

Shentsis, I., & Ben-Zvi, A. (1999). Within-season updating of seasonal forecast of Lake Kinneret inflow. *Journal of Hydrologic Engineering*, *4*(4), 381–385.

Shi, K., Zhang, Y., Qin, B., & Zhou, B. (2019). Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. *Science Bulletin*, *64*(20), 1540–1556.

Shuter, B. J., Schlesinger, D. A., & Zimmerman, A. P. (1983). Empirical predictors of annual surface water temperature cycles in North American lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *40*(10), 1838–1845.

Silveira Kupssinskü, L., Thomassim Guimarães, T., Menezes de Souza, E., C Zanotta, D., Roberto Veronez, M., Gonzaga, L., & Mauad, F. F. (2020). A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. *Sensors*, *20*(7), 2125.

Simis, S. G. H., Peters, S. W. M., & Gons, H. J. (2005). Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water. *Limnology and Oceanography*, *50*(1), 237–245.

Simis, S. G. H., Ruiz-Verdú, A., Domínguez-Gómez, J. A., Peña-Martinez, R., Peters, S. W. M., & Gons, H. J. (2007). Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass. *Remote Sensing of Environment*, *106*(4), 414–427.

Singh, S. P., & Singh, P. (2015). Effect of temperature and light on the growth of algae species: A review. *Renewable and Sustainable Energy Reviews*, *50*, 431–444.

Sinha, E., Michalak, M., & Balaji, V. (2017). Eutrophication will increase during the 21st century as a result of precipitation changes. *Science*, *357*(6349), 405–408. https://doi.org/10.1126/science.aan2409

Smith, S. R., Legler, D. M., Remigio, M. J., & O'Brien, J. J. (1999). Comparison of 1997–98 US temperature and precipitation anomalies to historical ENSO warm phases. *Journal of Climate*, *12*(12), 3507–3515.

Smith, T. M., Reynolds, R. W., Peterson, T. C., & Lawrimore, J. (2008). Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, *21*(10), 2283–2296.

Smith, V. H. (1982). The nitrogen and phosphorus dependence of algal biomass in lakes: An empirical and theoretical analysis 1. *Limnology and Oceanography*, *27*(6), 1101–1111.

Smith, V. H. (1985). PREDICTIVE MODELS FOR THE BIOMASS OF BLUE-GREEN ALGAE IN LAKES 1. *JAWRA Journal of the American Water Resources Association*, *21*(3), 433–439.

Smith, V. H. (2003). Eutrophication of freshwater and coastal marine ecosystems a global problem. *Environmental Science and Pollution Research*, *10*(2), 126–139. https://doi.org/10.1065/espr2002.12.142

Soley, C. (2016). Cyanobacteria Abundance Modeling: Development and Assessment of Season-Ahead Forecasts To Improve Beach Management on Lake Mendota.

Soranno, P A, Carpenter, S. R., & Lathrop, R. C. (1997). Internal phosphorus loading in Lake Mendota: response to external loads and weather. *Canadian Journal of Fisheries and Aquatic Sciences*, *54*(8), 1883–1893.

Soranno, Patricia A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., et al. (2017). LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. *GigaScience*, *6*(12). https://doi.org/10.1093/gigascience/gix101

Sòria-Perpinyà, X., Vicente, E., Urrego, P., Pereira-Sandoval, M., Ruíz-Verdú, A., Delegido, J., et al. (2020). Remote sensing of cyanobacterial blooms in a hypertrophic lagoon (Albufera of València, Eastern Iberian Peninsula) using multitemporal Sentinel-2 images. *Science of The Total Environment*, *698*, 134305. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.134305

Sòria-Perpinyà, X., Vicente, E., Urrego, P., Pereira-Sandoval, M., Tenjo, C., Ruíz-Verdú, A., et al. (2021). Validation of water quality monitoring algorithms for sentinel-2 and sentinel-3 in Mediterranean inland waters with in situ reflectance data. *Water*, *13*(5), 686.

Stevenson, S., Coats, S., Touma, D., Cole, J., Lehner, F., Fasullo, J., & Otto-Bliesner, B. (2022). Twenty-first century hydroclimate: A continually changing baseline, with more frequent extremes. *Proceedings of the National Academy of Sciences*, *119*(12), e2108124119.

Von Storch, H., & Zwiers, F. W. (2002). *Statistical analysis in climate research*. Cambridge university press.

Stow, C A, Carpenter, S. R., & Lathrop, B. C. (1997). A Bayesian observation error model to predict cyanobacterial biovolume from spring total phosphorus in Lake Mendota, Wisconsin. *Canadian Journal of Fisheries and Aquatic Sciences*, *54*(2), 464–473. https://doi.org/10.1139/f96-279

Stow, Craig A, Cha, Y., Johnson, L. T., Confesor, R., & Richards, R. P. (2015). Long-term and seasonal trend decomposition of Maumee River nutrient inputs to western Lake Erie. *Environmental Science & Technology*, *49*(6), 3392–3400.

Stumpf, R. P., Wynne, T. T., Baker, D. B., & Fahnenstiel, G. L. (2012). Interannual variability of cyanobacterial blooms in Lake Erie.

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., et al. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae*, *54*, 160–173.

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

Su, H., Lu, X., Chen, Z., Zhang, H., Lu, W., & Wu, W. (2021). Estimating coastal chlorophyll-a concentration from time-series OLCI data based on machine learning. *Remote Sensing*, *13*(4), 576.

Sunda, W. G., Graneli, E., & Gobler, C. J. (2006). Positive feedback and the development and persistence of ecosystem disruptive algal blooms 1. *Journal of Phycology*, *42*(5), 963–974.

Survey, U. S. G. (2021a). National Water Information System data available on the World Wide Web (USGS Water Data for the Nation). Retrieved September 22, 2021, from https://waterdata.usgs.gov/usa/nwis/uv?site_no=05427718

Survey, U. S. G. (2021b). National Water Information System data available on the World Wide Web (USGS Water Data for the Nation). Retrieved September 22, 2021, from https://waterdata.usgs.gov/nwis/uv?site_no=05427850

Swingle, H. S. (1968). Fish kills caused by phytoplankton blooms and their prevention. *FAO Fish Rep*, *44*(5), 407–411.

Taranu, Z. E., Zurawell, R. W., Pick, F., & Gregory-Eaves, I. (2012). Predicting cyanobacterial dynamics in the face of global change: the importance of scale and environmental context. *Global Change Biology*, *18*(12), 3477–3490.

Taranu, Z. E., Gregory-Eaves, I., Leavitt, P. R., Bunting, L., Buchaca, T., Catalan, J., et al. (2015). Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. *Ecology Letters*, *18*(4), 375–384.

Teklehaimanot, H. D., Lipsitch, M., Teklehaimanot, A., & Schwartz, J. (2004). Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. *Malaria Journal*, *3*(1), 1–11.

Tian, S., Guo, H., Xu, W., Zhu, X., Wang, B., Zeng, Q., et al. (2023). Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. *Environmental Science and Pollution Research*, *30*(7), 18617–18630.

Tompkins, A. M., Lowe, R., Nissan, H., Martiny, N., Roucou, P., Thomson, M. C., & Nakazawa, T. (2019). Predicting climate impacts on health at sub-seasonal to seasonal timescales. In *Sub-Seasonal to Seasonal Prediction* (pp. 455–477). Elsevier.

Tootle, G. A., Piechota, T. C., & Singh, A. (2005). Coupled oceanic-atmospheric variability and US streamflow. *Water Resources Research*, *41*(12).

Towler, E., Rajagopalan, B., Gilleland, E., Summers, R. S., Yates, D., & Katz, R. W. (2010). Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resources Research*, *46*(11).

Trenberth, K. E., & Caron, J. M. (2000). The Southern Oscillation revisited: Sea level pressures, surface temperatures, and precipitation. *Journal of Climate*, *13*(24), 4358–4365.

Trombetta, T., Vidussi, F., Roques, C., Mas, S., Scotti, M., & Mostajir, B. (2021). Co-occurrence networks reveal the central role of temperature in structuring the plankton community of the Thau Lagoon. *Scientific Reports*, *11*(1), 1–14.

Umaña, J. D., Montenegro-Torres, J., & Otero, J. (2024). epiCo: provides statistical and visualization tools for the analysis of outbreaks of vector-borne diseases (VBDs) in Colombia.

U.S. Geological Survey., 2020b. Landsat-5 Surface Reflectance Tier 1. Accessed: 2020-09-16. Retrieved from https://code.earthengine.google.com.

Vanhellemont, Q., & Ruddick, K. (2018). Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. *Remote Sensing of Environment*, *216*, 586–597.

Vanhellemont, Q., & Ruddick, K. (2021). Atmospheric correction of Sentinel-3/OLCI data for mapping of suspended particulate matter and chlorophyll-a concentration in Belgian turbid coastal waters. *Remote Sensing of Environment*, *256*, 112284.

Vanni, M. J., & Temte, J. (1990). Seasonal patterns of grazing and nutrient limitation of phytoplankton in a eutrophic lake. *Limnology and Oceanography*, *35*(3), 697–709.

Vanni, M. J., Luecke, C., Kitchell, J. F., & Magnuson, J. J. (1990). Effects of planktivorous fish mass mortality on the plankton community of Lake Mendota, Wisconsin: implications for biomanipulation. In *Biomanipulation Tool for Water Management: Proceedings of an International Conference held in Amsterdam, The Netherlands, 8–11 August, 1989* (pp. 329–336). Springer.

Villar, L. A., Rojas, D. P., Besada-Lombana, S., & Sarti, E. (2015). Epidemiological trends of dengue disease in Colombia (2000-2011): a systematic review. *PLoS Neglected Tropical Diseases*, *9*(3), e0003499.

Villegas, J. G., Gutiérrez, E. V., Barrera Ferro, D., Muriel, O., Felipe, A., Paredes Bayona, J. E., et al. (2020). *Aplicaciones de investigación de operaciones en sistemas de salud en Colombia*. Editorial Pontificia Universidad Javeriana.

Vincent, R. K., Qin, X., McKay, R. M. L., Miner, J., Czajkowski, K., Savino, J., & Bridgeman, T. (2004). Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial blooms in Lake Erie. *Remote Sensing of Environment*, *89*(3), 381–392.

Visbeck, M. H., Hurrell, J. W., Polvani, L., & Cullen, H. M. (2001). The North Atlantic Oscillation: past, present, and future. *Proceedings of the National Academy of Sciences*, *98*(23), 12876–12877.

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1889–1899.

Vitart, F., & Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Npj Climate and Atmospheric Science*, *1*(1), 3.

Vitart, F., Robertson, A. W., & Anderson, D. L. T. (2012). Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization*, *61*(2), 23.

Vollenweider, R. A. (1971). *Scientific fundamentals of the eutrophication of lakes and flowing waters, with particular reference to nitrogen and phosphorus as factors in eutrophication*. Organisation for economic co-operation and development Paris.

Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, *15*(1), 1625–1651.

Wagner, C., & Adrian, R. (2009). Cyanobacteria dominance: quantifying the effects of climate change. *Limnology and Oceanography*, *54*(6part2), 2460–2468.

Walsh, J. R., Lathrop, R. C., & Vander Zanden, M. J. (2017). Invasive invertebrate predator, Bythotrephes longimanus, reverses trophic cascade in a north-temperate lake. *Limnology and Oceanography*, *62*(6), 2498–2509.

Waylen, P., & Poveda, G. (2002). El Niño–Southern Oscillation and aspects of western South American hydro-climatology. *Hydrological Processes*, *16*(6), 1247–1260.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., et al. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, *24*(3), 315–325.

Whiting, E. C., Khan, A., & Gubler, W. D. (2001). Effect of temperature and water potential on survival and mycelial growth of Phaeomoniella chlamydospora and Phaeoacremonium spp. *Plant Disease*, *85*(2), 195–201.

WHO. (2019). Dengue vaccine: WHO position paper, September 2018 – Recommendations. *Vaccine*, *37*(35), 4848–4849. https://doi.org/https://doi.org/10.1016/j.vaccine.2018.09.063

Wilkinson, G. M., Walter, J. A., Buelo, C. D., & Pace, M. L. (2021). No evidence of widespread algal bloom intensification in hundreds of lakes. *Frontiers in Ecology and the Environment*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/fee.2421

Wimberly, M. C., Nekorchuk, D. M., & Kankanala, R. R. (2022). Cloud-based applications for accessing satellite Earth observations to support malaria early warning. *Scientific Data*, *9*(1), 1–11.

Wines, M. (2014, August 4). Behind Toledo's Water Crisis, a Long-Troubled Lake Erie. *New York Times*. Retrieved from https://www.nytimes.com/2014/08/05/us/lifting-ban-toledo-says-its-water-is-safe-to-drink-again.html

Wood, A. W., Kumar, A., & Lettenmaier, D. P. (2005). A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States. *Journal of Geophysical Research: Atmospheres*, *110*(D4).

Woolway, R. I., Kraemer, B. M., Lenters, J. D., Merchant, C. J., O'Reilly, C. M., & Sharma, S. (2020). Global lake responses to climate change. *Nature Reviews Earth & Environment*, *1*(8), 388–403.

Wuertz, D., Lawrimore, J., & Korzeniewski, B. (2018). Cooperative observer program (COOP) hourly precipitation data (HPD), version 2.0. NOAA National Centers for Environmental Information.

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

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

Wynne, Timothy T, Stumpf, R. P., Tomlinson, M. C., Fahnenstiel, G. L., Dyble, J., Schwab, D. J., & Joshi, S. J. (2013). Evolution of a cyanobacterial bloom forecast system in western Lake Erie: development and initial evaluation. *Journal of Great Lakes Research*, *39*, 90–99.

Xiao, X., He, J., Yu, Y., Cazelles, B., Li, M., Jiang, Q., & Xu, C. (2019). Teleconnection between phytoplankton dynamics in north temperate lakes and global climatic oscillation by time-frequency analysis. *Water Research*, *154*, 267–276. https://doi.org/https://doi.org/10.1016/j.watres.2019.01.056

Yan, Y., Bao, Z., & Shao, J. (2018). Phycocyanin concentration retrieval in inland waters: A comparative review of the remote sensing techniques and algorithms. *Journal of Great Lakes Research*, *44*(4), 748–755.

Yuan, H.-Y., Liang, J., Lin, P.-S., Sucipto, K., Tsegaye, M. M., Wen, T.-H., et al. (2020). The effects of seasonal climate variability on dengue annual incidence in Hong Kong: A modelling study. *Scientific Reports*, *10*(1), 4297.

Zhang, H., Hu, W., Gu, K., Li, Q., Zheng, D., & Zhai, S. (2013). An improved ecological model and software for short-term algal bloom forecasting. *Environmental Modelling & Software*, *48*, 152–162.

Zhang, M., Duan, H., Shi, X., Yu, Y., & Kong, F. (2012). Contributions of meteorology to the phenology of cyanobacterial blooms: implications for future climate change. *Water Research*, *46*(2), 442–452.

Zhang, M., Qin, B., Yu, Y., Yang, Z., Shi, X., & Kong, F. (2016). Effects of temperature fluctuation on the development of cyanobacterial dominance in spring: implication of future climate change. *Hydrobiologia*, *763*(1), 135–146.

Zhu, B., Cao, H., Li, G., Du, W., Xu, G., Santo Domingo, J., et al. (2019). Biodiversity and dynamics of cyanobacterial communities during blooms in temperate lake (Harsha Lake, Ohio, USA). *Harmful Algae*, *82*, 9–18.

Zhu, S., Ptak, M., Yaseen, Z. M., Dai, J., & Sivakumar, B. (2020). Forecasting surface water temperature in lakes: A comparison of approaches. *Journal of Hydrology*, *585*, 124809.

Zimmerman, B. G., Vimont, D. J., & Block, P. J. (2016). Utilizing the state of ENSO as a means for season-ahead predictor selection. *Water Resources Research*, *52*(5), 3761–3774.