Identification of Satellite Indicators for Predicting CyanoHABs in Kansas

By

Cheyenne A. Hillman
B.S., University of Kansas, 2021

Submitted to the graduate degree program in Civil, Environmental, and Architectural Engineering and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science.

Chair: Joshua Roundy, Ph.D.

Admin Husic, Ph.D.

Amy Hansen, Ph.D.

Ted Harris, Ph.D.

Date Defended: 3 June 2022

The thesis committee for Cheyenne A Hillman certifies that this is the approved version of the following thesis:

Identification of Satellite Indicators for Predicting CyanoHABs in Kansas

Chair: Joshua Roundy, Ph.D.

Admin Husic, Ph.D.

Amy Hansen, Ph.D.

Ted Harris, Ph.D.

Date Approved: 9 June 2022

Abstract

Cyanobacterial Harmful Algal Blooms (HABs) degrade water quality by producing harmful toxins and causing significant diel changes in water column pH and dissolved oxygen concentrations, leading to degradation of water quality that disrupts food webs, and has negative ecological, social, recreational, and economic impacts. In Kansas, the ubiquity and impact of HABs necessitates large-scale monitoring and prediction of the events, however, no clear indicator or predictor for HABs has been established. The goal of this study is to determine if surface observations from satellite retrievals contain information about the development of HAB events and if this information can be assimilated into a 1-D lake model to improve the prediction of HAB events in Kansas. Different environmental variables are explored to determine their suitability as predictors or indicators of HABs, and with identified candidates, nonlinear regression and regression tree models are created at Cheney Reservoirs with cyanobacteria data. To evaluate their transferability, they are then tested with sediment core pigment data primarily at Marion Reservoir. These results showed that MODIS land surface temperature satellite data paired with NLDAS precipitation, windspeed, and shortwave radiation gave the most promising results for application at both Cheney and Marion. These results can be used for future assimilation into lake models to help with better prediction and modeling of blooms.

Acknowledgements

I want to thank my advisor, Dr. Joshua Roundy, for his guidance through both my undergraduate and graduate studies. The coursework I have taken with him helped drive me towards pursing a graduate education to further develop my knowledge in hydraulics and hydrology. His dedication to helping students and commitment to research has been invaluable, and I could not have imagined having a better advisor and mentor throughout my time at the University of Kansas.

Thank you to Dr. Admin Husic for introducing me to research during my undergraduate degree, and whose classes helped me dig into my interests in the field. I am also grateful to Dr. Amy Hansen for being an excellent resource in my coursework and for serving on my committee. I would also like to thank Dr. Ted Harris, who has provided valuable advice and resources throughout my research.

I am extremely grateful to my husband for his love and unfailing support throughout school and through the process of researching and writing my thesis. Thank you for supporting me through everything and giving help and advice along the way. Last, I would like to thank my mother who has always been my biggest supporter. Her belief in me and encouragement has kept my motivation high throughout this process.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction

The transportation and storage of water in its various forms is what makes this planet unique. Humans in particular rely on the water cycle and its complex interactions with ecosystems that support life and lifestyle. Within the water cycle, surface water is an important resource that supports human society as well as many ecological systems and provides an important habitat for many species (US EPA 2017). These same vital ecosystems also provide multi-use resources for human societies. Specifically, many lakes and reservoirs provide drinking water, irrigation for agriculture, recreation, and more. Protecting surface water resources is essential for maintaining their current uses and protecting the well-being of all species that rely on these vital systems. However, human induced changes from urban development, reservoir construction, agricultural practices, and increased use of fossil fuels leading to climate change have greatly impacted surface water ecosystems (Brikowski 2008; Wang and Hejazi 2011; NOAA National Centers for Environmental Information 2020).

In the National Oceanic and Atmospheric Administration's (NOAA) 2020 State of the Climate report, they reported increased greenhouse gases are leading to higher temperatures, with the global temperature in 2020 being 0.98 °C higher than prior averages and being the second warmest year in the past 141 years (NOAA National Centers for Environmental Information 2020). The report also found that North America experienced even warmer temperatures overall, being 1.16 °C above the prior centuries average. In Kansas, where the western and eastern sides of the state have large climate variability, there has been a statewide trend shown towards warming 0.06 °C per year from 1985 to 2015 (Lin et al. 2017). Warming temperatures also leads to intensification of the hydrologic cycle, including more atmospheric water vapor, changes in cloud properties, and

changes in snow cover (Huntington 2006). Changes in these aspects of the hydrologic cycle cause a positive feedback cycle that continues to escalate warming. In addition to temperature changes, precipitation has changed in the state was well, with the western portion of the state becoming drier, and the eastern side potentially getting wetter (Lin et al. 2017). Precipitation in the U.S. has been the reason for most of the variability in runoff over the past century, having a greater effect than temperature (McCabe and Wolock 2011). Land use change may also impact the amount of runoff, as increased impervious surfaces allow for more runoff to flow across the land surface. While this is a problem, climate change has been attributed to have a greater impact in some areas that land use change (Shang et al. 2019). The impacts of changing temperatures can have devastating impacts on water resources. In the drier western side of the state, four Federal reservoirs have had negative water balances due to declines in streamflow and increases in evaporation correlated with climate change (Brikowski 2008). For these reservoirs, their water resources are expected to have over a 50% decline from 2007 to 2050. The impacts on water resources can lead to further economic damage. In western Kansas, the supply of groundwater in the Ogallala aquifer has declined due to high demands for irrigation, and mean annual streamflow has been shown to be declining primarily due to human activities (Araya et al. 2017; Wang and Hejazi 2011). In western Kansas, maize yield is expected to be reduced by up to 46% by 2050 (Araya et al. 2017). The primary driver of these declines is a shortened growing period caused by higher temperatures. With climate change worsening over time, the water resources in Kansas are at a greater risk for depletion.

While human induced climate change is a great threat to surface waters, other environmental changes due to anthropogenic influences have caused increased eutrophication, especially increased amount of nitrogen and phosphorous in water (Paerl and Huisman 2009).

Human-related influences include nutrient loading from agricultural, urban, and industrial areas. The increased eutrophication has degraded water quality through increased frequency and intensity of cyanobacteria harmful algal blooms (cyanoHABs). CyanoHABs disrupt food webs and have negative ecological, social, recreational, and economic impacts (Paerl and Otten 2012; Anderson et al. 2000). These impacts include harming plants and subsequently the habitats of aquatic life, depleting oxygen in the water column when the blooms die off, closing down recreational areas, increasing the cost of monitoring and management, and increasing cost of treatment for public health impacts.

The Midwestern United States is particularly plagued by CyanoHABs, due to extensive nutrient loading (Downing, Watson, and McCauley 2001; Graham et al. 2004; 2010). In Kansas, over 50 lakes have experienced CyanoHABs in the last 10 years. During 2010 to 2017, Milford Lake experienced HAB events on 41% of days, and in 2011, cyanotoxins exceeded the World Health Organization's threshold for no effects by 7,500 times, making it the highest toxin concentration on record (Graham et al. 2012; KDHE n.d.) The increased occurrence of CyanoHABs and their harmful effects in Kansas has made it a top priority for the Kansas Water Office (KWO) and the Kansas Department of Health and the Environment (KDHE). Given the ubiquity and impact of CyanoHABs in Kansas, there is a need for greater understanding and prediction of CyanoHAB events. KDHE defines algal blooms as "a dense growth of any type of algae" and harmful algal blooms as "a dense growth of algae that has the potential for creating toxins or other nuisance compounds" (KDHE n.d.).

One of the main limitations of creating a predictive model for CyanoHABs in Kansas is the lack of available data on when and where CyanoHABs have occurred. Furthermore, the ubiquity and impact of CyanoHABs in Kansas necessitates the creation of robust monitoring and

prediction framework that is applicable on a statewide scale. One possible pathway to overcome these limitations is to utilize existing data of algae in conjunction with satellite remote sensing data to create predictive models of CyanoHABs. However, this requires that satellite data are able to capture key environmental conditions that facilitate the development and continuation of cyanobacteria within Kansas reservoirs.

1.2 CyanoHAB Environmental Conditions

The rise in global temperatures has the potential to increase CyanoHABs due to the fact that Cyanobacteria grow best at around 25 °C (Paerl and Otten 2012). On top of this, the increases in temperature can lead to more vertical stratification of temperature in the water column (Paerl and Huisman 2008; 2009). Cyanobacteria are also more prolific in stratified conditions, especially in low-wind and warm conditions that allow them to float at the surface and accumulate (Paerl and Huisman 2009). Climate change also leads to other conditions beneficial to cyanobacteria growth. Intense precipitation can lead to higher nutrient transport into water bodies (Paerl and Huisman 2008). In the short-term, blooms may be diluted and washed away by precipitation events, but in the longer-term, nutrient loads that entered the reservoir during these events can prompt later blooms (Paerl and Huisman 2008). Drought may also lead to cyanobacteria dominance by increasing salinity of freshwater, as cyanobacteria can be more tolerant to high salinities than other freshwater phytoplankton (Paerl and Huisman 2008; Tonk et al. 2007). The overall changes in the climate are creating an environment that is much more beneficial for cyanobacteria growth and dominance in freshwaters.

Large growth of CyanoHABs makes eutrophic waters more turbid, which leads to poor conditions for aquatic macrophytes to thrive, which provide habitats for other species (Paerl and Huisman 2009)**.** On top of this, decomposition of cyanobacteria consumes large amounts of

dissolved oxygen when blooms begin to die off and can lead to depletion of dissolved oxygen at night, killing fish (Paerl and Huisman 2009). Cyanobacteria can also produce toxins, threatening the safe use of water for drinking, recreation, fishing, and irrigation (Carmichael 2001). Health problems from toxins can occur in both humans and animals, and include diseases in the liver, digestive system, and skin, and even neurological issues or death (Carmichael 2001; Huisman, Matthijs, and Visser 2005). These toxins can also build up through the food chain, reaching human food sources (Huppert et al. 2005). Cyanobacteria can also cause more aesthetic issues with drinking water supplies. They can produce taste-and-odor compounds, making the water unpleasant for smell and taste (Graham et al. 2010). While these are not necessarily associated with health risks, they require modified drinking water treatment so funding must be put into monitoring blooms be able to make appropriate adjustments (Graham et al. 2010; Anderson et al. 2000)

At Cheney Reservoir in Kansas, cyanobacteria have been shown to be an episodic water-quality issue with clear seasonal patterns (Graham, Foster, and Kramer 2017). Their presence has been shown to have a relationship with increased nutrient and decreased sediment concentrations (Graham, Foster, and Kramer 2017). The main limitation of growth is likely decreased available light from sediment concentrations (Christensen et al. 2006). The climatic conditions and their influence on the environment of cyanobacteria helps set up the use of remote sensing for the detection and prediction of CyanoHABs in Kansas.

1.3 Remote Sensing

Remote sensing allows for indirect monitoring of the land surface environment using sensors onboard satellites and aircraft (“Remote Sensing: An Overview" n.d.). Satellite sensors are able to detect reflected or emitted energy, and this data is then processed and used to understand

and monitor surface characteristics. Satellite data has been used for detecting land cover changes, mapping deforestation, monitoring biodiversity loss, and much more (Wulder and Coops 2014). It has also been used to monitor ocean and inland water quality (McClain 2009; Mouw et al. 2015).

Remote sensing has been used for detection of HAB events as far back as 1972, with LandSat 1 (Murphy et al. 1975). Since then, many agencies have been able to use satellite water quality data to detect and proactively close water bodies to ensure safety. Reactive management strategies where events are detected through satellite data and then teams are sent out to validate the measurements are practices by many states (Stroming et al. 2020). Using satellite remote sensing can allow for more proactive strategies in CyanoHAB monitoring and detection, especially in conjunction with in situ sensors that provide continuous monitoring (Stroming et al. 2020). With these factors in mind, the incorporation of satellite data in monitoring strategies has been demonstrated and quantified to have socioeconomic benefits and can help avoid negative human health outcomes (Stroming et al. 2020; Papenfus et al. 2020; Coffer et al. 2020).

Several efforts to use satellite data to monitor and/or predict blooms have been undertaken. Project CyAN, led by the National Aeronautics and Space Administration (NASA), NOAA, the Environmental Protection Agency (EPA), and the U.S. Geological Survey (USGS), uses the Medium Resolution Imaging Spectrometer (MERIS) to develop early warning indicators for cyanobacteria blooms across the contiguous U.S. (US EPA 2014). NOAA also has another monitoring program, the "Experimental Lake Erie Harmful Algal Bloom Bulletin," which has a goal of monitoring and forecasting HABs in a portion of Lake Erie using satellite images (Stumpf and Dupuy 2016). The Kansas Biological Survey (KBS) has previously investigated the relationship with the normalized difference vegetation index (NDVI) at some Kansas reservoirs where NDVI was successfully used to qualitatively represent greenness (Dzialowski et al. 2008).

1.4 Hypothesis and Research Questions

While using satellite data to monitor HAB events has been widely applied, predicting HAB events in Kansas through satellite data is not established due to several key limitations:

- Although major factors are clear (nutrients, climate warming), the proximal drivers of HAB events are not well understood due to a lack of measurements of CyanoHABs across the state of Kansas that could be used to create predictive models.
- There is no clear connection between satellite measurements and lake models needed in order to use satellite data to improve predictions of HAB events.

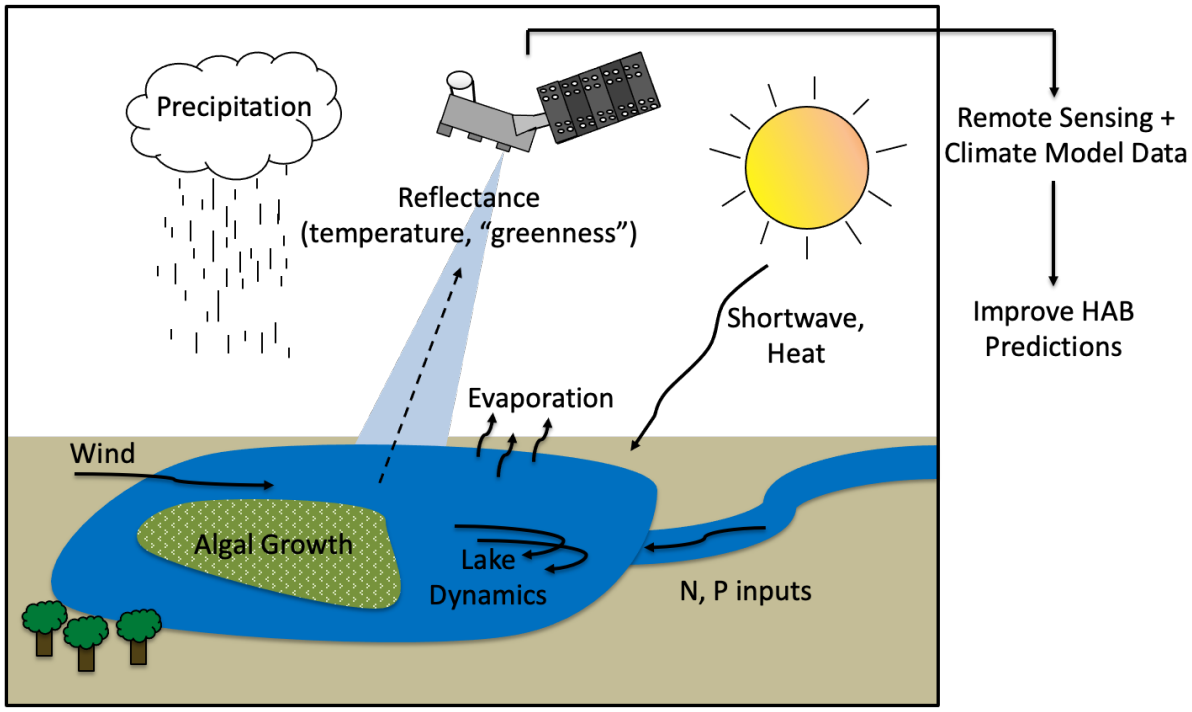

Figure 1. Conceptual diagram of the use of remote sensing for improving the prediction CyanoHABs in Kansas and paving the way for assimilating satellite data into lake models.

The objective of this work is to overcome these limitations by using satellite data to establish key predictive drivers of HAB events and use these relationships to improve lake water quality models that can be used for the prediction of HAB events. I hypothesize that satellite images contain key information about the formation and development of HAB events and that statistical models can

be created that provide predictive power for CyanoHAB events. This hypothesis will be tested by first identifying the key information from satellite data. Potential satellite indicators (temperature and "greenness") are combined with other hydroclimate variables (shortwave radiation, precipitation, and wind speed) to explore their relationship with cyanobacteria growth in reservoirs (Figure 1). The key research questions that will be addressed in this work are:

1. What are the trends and relationships between environmental variables at Kansas reservoirs?
2. What drivers are key to prediction of HAB events in Kansas reservoirs?
3. What model framework provides the best predictions and also provides transferability to other locations?

Answering these research questions will identify key relationships between satellite indicators and algae blooms that will help provide a basis for ultimately assimilating satellite data into lake models for better prediction of HAB events that can be used in earth system modeling.

Chapter 2: Datasets & Methods

2.1 Study Locations

To evaluate indicators as potential sources of predictability for CyanoHABs, data was analyzed over all major Kansas reservoirs. Big Hill Reservoir and Keith Sebelius Lake were initially included in the dataset, but due to their small size there were no pixels in some of the remote sensing imagery classified as "open water" and were therefore omitted from this analysis. Table 1 provides a list of reservoirs used in this analysis. While all lakes shown in Table 1 were analyzed for trends and correlations using a subset of environmental variables from remote sensing, reanalysis, and in-situ measurements, much of the predictability analysis focuses on Cheney and Marion due to their similarities and availability of key in-situ measurements need to create the prediction models.

Table 1. Reservoirs for Analysis

Name	Area (mi^2)
Cedar Bluff	10.73
Cheney	14.92
Clinton	10.94
Council Grove	5.06
El Dorado	12.50
Elk City	7.03
Fall River	3.82
Hillsdale	7.16
John Redmond	14.69
Kanopolis	5.32
Kirwin	7.94
Lovewell	4.67
Marion	9.63
Melvern	10.8
Milford	24.71
Perry	17.19
Pomona	6.25
Toronto	4.38
Tuttle Creek	19.31
Waconda	19.69
Webster	5.89
Wilson	14.13

2.2 Datasets

To achieve the goals of the project, several indicators were explored for their potential for predicting CyanoHABs. Potential indicators include air and surface temperature, windspeed, shortwave radiation and surface reflectance indices that capture the greenness of the surface such as the enhanced vegetation index (EVI), NDVI, and the floating algae index (FAI). Each of these indicators was selected as it has the potential to capture a key aspect of environment that promotes blooms. Temperature, wind and shortwave radiation were chosen since blooms are known to thrive in warm temperatures, low-wind (reduced lake mixing) and with sufficient light to facilitate

photosynthesis and continued growth. The EVI, NDVI, and FAI were chosen due to their potential ability to detect the color of blooms from remotely sensed data. Furthermore, identifying predictive relationships also requires in-situ measurements of CyanoHABs in order to train and validate a predictive model. All of the datasets used in this analysis are broken down into three types, in-situ, remotely sensed and reanalysis are each discussed in more detail below.

2.2.1 In-Situ Datasets

The first set of in-situ data consist of environmental variables measured by the U.S. Army Corps of Engineers and the National Weather Service at or near the reservoir locations. For this analysis, these data sets were primarily used to compare with other data sets that provide a broader spatial coverage. The U.S. Army Corps of Engineers (USACE) Tulsa District provides monitoring data for the following reservoirs: Cheney, Council Grove, El Dorado, Elk City, Fall River, John Redmond, and Toronto. Cheney Reservoir's data is not included as it does not contain the necessary information (temperature) for the analysis. The Kansas City District of USACE oversees the remaining reservoirs in the northern portion of Kansas. Publicly available monitoring data is not available through USACE for the locations in the Kansas City District. Datasets for the Tulsa District's locations are publicly available on the USACE Tulsa District Water Control Data System webpage (https://www.swt-wc.usace.army.mil). This dataset includes measurements on the hour of precipitation, elevation, storage, inflow, outflow, air temperature, wind direction, wind speed, relative humidity, and shortwave radiation at the dam for each reservoir. All datasets begin in April 2017 (varying start days) to July 31, 2021. All these datasets have four-month gaps at the end of 2018, and several other smaller gaps throughout the dataset. Data was often presented in incorrect columns, though these website formatting issues were identified and corrected for the final compiled dataset. NOAA monitoring data was collected for several locations where NOAA

stations could be identified within approximately 30 miles of the reservoir and included windspeed information. Data was collected through NOAA's Climate Data Online tool (https://www.ncdc.noaa.gov/cdo-web/). NOAA monitoring information was collected for January 1, 2002 to present and consisted of daily averages. The key piece of information provided by these datasets was the windspeed, and it was mainly used in analysis for Cheney Reservoir to compare with other sources of wind speed. The NOAA station closest to Cheney Reservoir is located at Eisenhower Airport in Wichita, Kansas, approximately 25 miles away from the USGS station on the dam.

The main data set used to create and validate the CyanoHAB prediction model is from a high frequency sensor at Cheney Reservoir maintained by the USGS. The USGS station is located near the dam and the sensor measures fluorescence and is then converted and made available in cyanobacteria fluorescence of phycocyanin (fPC) in cells per milliliter at hourly intervals from October 1, 2012 until it was discontinued on March 12, 2015 (U.S. Geological Survey 2022). From October 1, 2014 to present, a different sensor was installed at Cheney. This sensor also measures fluorescence but is reported in relative fluorescence units (RFU) in 30-minute intervals, which is used to qualitatively measure the phycocyanin. Figure 2 shows the relationship between actual samples of cyanobacteria cell count and the sensor output to test the relationship between sensor measurements and actual cell count values. These values are not the same, but they have a positive correlation that the higher sensor values are generally indicative of higher cell concentration and

should capture bloom events. These data sets provide the basis to create and test the relationship between cyanobacteria and the potential indicators.

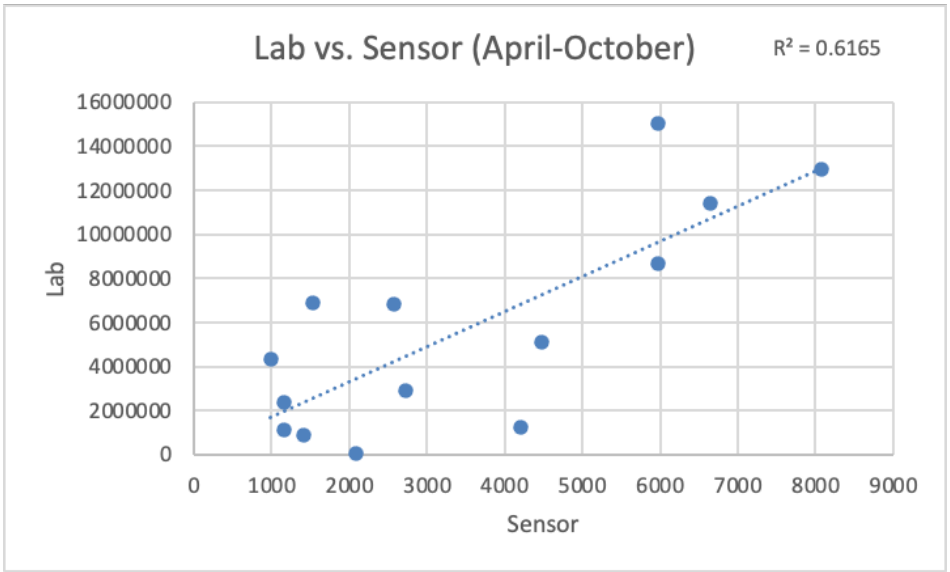

Figure 2. Comparison of lab measurements of cyanobacteria cell count to sensor measurements. Sensor measurements include the period of interest, April through October, and the sensor measurement is taken at the nearest time to the time the lab sample was taken.

Another data set used to validate the prediction models for CyanoHABs is the KBS sediment core data sets. These sediment core data sets are available for 6 reservoirs: Marion, Milford, Lovewell, Kanopolis, Perry, and Webster. Each of these datasets date back to their respective reservoir's date of impoundment (Harris et al. 2020b; 2020a; 2021). To collect samples, a core tube was hammered into the reservoir sediments deep enough to reach the pre-impoundment soil at locations denoted as S1 (sample site 1) and S2 (sample site 2) shown in Figure 3 (Harris et al. 2020b). Samples were transported to and assessed at the KBS lab. The cores were tested in 1 cm intervals. Plutonium nuclides were utilized to date the increments, where peaks in the nuclide are representative of sediment from before impoundment. Phytoplankton pigments were also analyzed for the samples, which provides a timeline of dead phytoplankton buildup in the sediment. Pigments of interest for identifying cyanobacteria include echineone, zeaxanthin-lutein, and canthaxanthin. Echineone is unique to cyanobacteria. Zeaxanthin-lutein is present in

cyanobacteria but also in other green algae, and canthaxanthin is possibly indicative of nitrogen-fixing cyanobacteria. The methods in Leavitt & Hodgson were used to assess phytoplankton pigments (2002). For this analysis, we only used the sediment core sample at sample site 2 at each

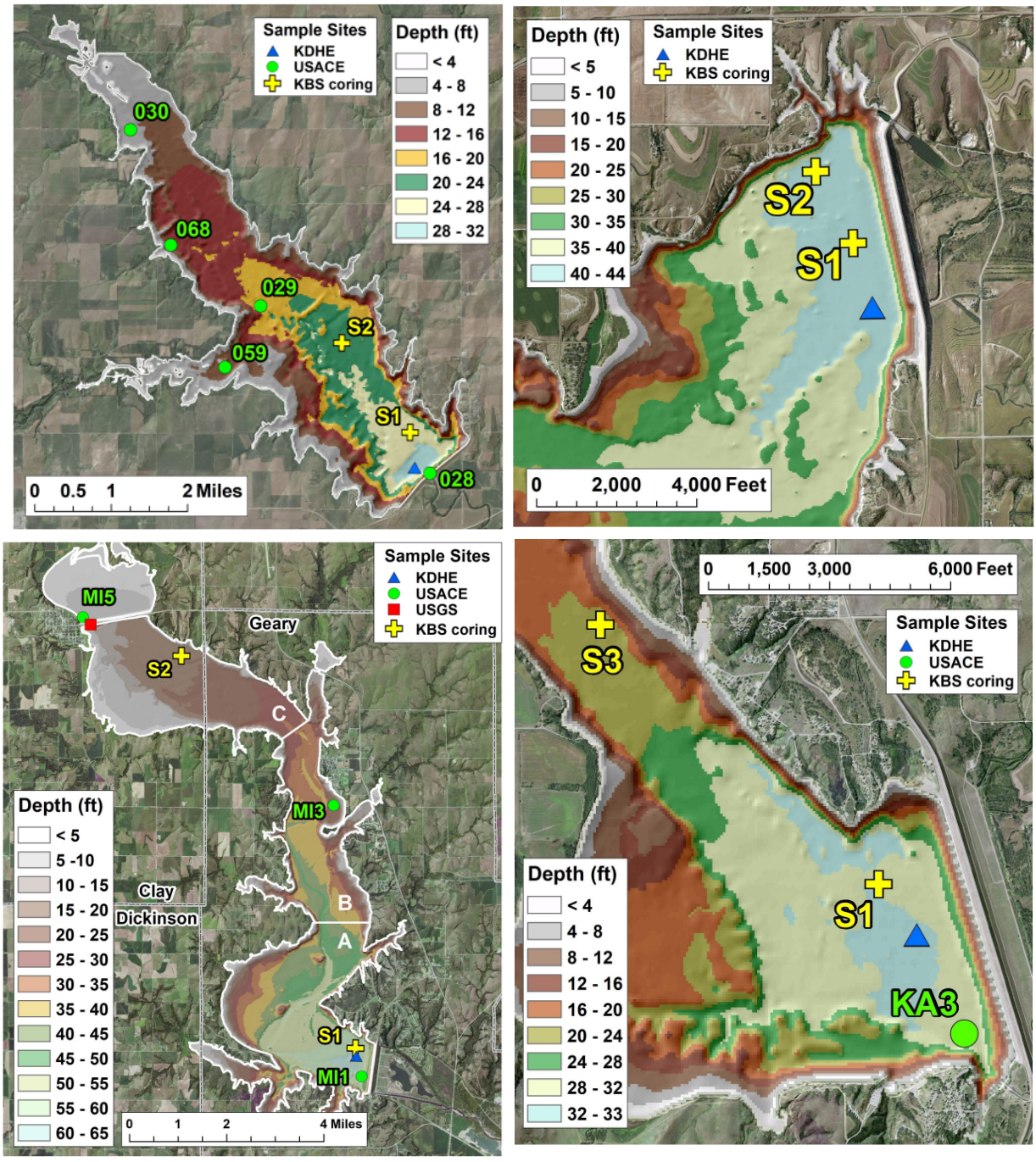

Figure 3. Bathymetric maps of Marion Reservoir (top left), Webster Reservoir (top right), Milford Reservoir (bottom left), and Kanopolis Reservoir (bottom right) with sediment core locations (Images from: Harris et al. 2020b; 2020a; 2021)

reservoir as these sites were located closer to the dam and are thought to be more representative of where blooms would occur. While the sediment core dataset provides a long record dating back to impoundment, it is also temporally coarse, only providing annual concentration of each pigment with some years missing due to the uncertainty in dating pigments in the sediment. Sediment cores are primarily useful because they can be isolated and identified after vegetative cell structures are degraded and, when sediment mixing is relatively low, they provide a layered timeline of dead phytoplankton accumulation (Harris et al. 2020b). There is also a large amount of uncertainty in the dating process itself, so the dates themselves may not be reliable. Yet despite these uncertainties, the sediment core provides an independent data set for validating the prediction models.

2.2.2 Remote Sensing Datasets

In this analysis, two remote sensing data sets were analyzed NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) and the USGS LandSat Collection 1. For both data sets, surface temperature and surface reflectance indices are analyzed, however, each of these data sets have a different temporal and spatial resolution due to sensors and orbit of each of the satellites. MODIS is a sensor that collects earth and climate measurements through remote sensing. One of the main advantages of using the MODIS satellite data set is its ease-of-access through NASA's Global Subsets Tool where products can be selected for multiple coordinates at once and then easily downloaded. Many reservoirs were large enough to require multiple data selections, and these datasets were combined during later processing. Moreover, the system provides downloadable spreadsheets of the data for each pixel, meaning no additional image processing is required to retrieve the necessary information. Several different datasets from MODIS Terra and Aqua, the two satellites in the program, are included. Terra's temporal extent is from February 18,

2000, to present, and Aqua's is from July 4, 2002 to present. The datasets were analyzed for 22 Kanas reservoirs (Table 1). MOD11A2 (Terra) and MYD11A2 (Aqua) provide the 8-day night and day land surface temperature (LST) at a 1 km spatial resolution (Wan, Hook, and Hulley 2015a; 2015b). MOD13Q1 (Terra) and MYD13A1 (Aqua) provide the 16-day EVI and NDVI at a 250 m spatial resolution (K. Didan 2015a; 2015b). MOD09A1 (Terra) and MYD09A1 (Aqua) provide the 8-day surface spectral reflectance's at a 500 m spatial resolution (Vermonte 2015a; 2015b). The available surface reflectance bands were also corrected for atmospheric conditions for use in calculating other reflectance indices such as the floating algae index. For each lake the average surface temperature and reflectance indices was calculated over all open water grids. As a reference, the 1-km grid resolution of the MODIS data resulted in 32 grid cells over Chaney reservoir.

The USGS LandSat Collection 1 U.S. Analysis Ready Data (ARD), which includes data combined from multiple LandSat missions, was also used as a source for land surface temperature and surface reflectance. As compared to the MODIS dataset, the LandSat datasets required substantial more processing, but provide a higher spatial resolution and sometimes a higher temporal resolution (Dwyer et al. 2018). LandSat 1, 2, and 3 have a spatial resolution of 60 meters, and LandSat 4, 5, 7, 8, and 9 have a spatial resolution of 30 meters and overpass timing of approximately 16 days. In order to extract the needed surface temperature and surface reflectance from the data set, each LandSat image for the ARD grid had to be downloaded, checked for cloud cover and swath coverage over the lake. This process was automated using a Python script that utilized the USGS API to request and download thousands of images. The ARD images have a scene size of 170 km by 185 km and for each image the lake area was identified, and the average surface temperature and calculated vegetation index was extracted. Because the LandSat dataset

includes multiple satellites, partial swath coverage of the lake area, and the possibility of cloud coverage, it does not have a consistent temporal resolution. As a reference, Cheney reservoir consists of 38,459 grid cells in the LandSat data.

2.2.3 Reanalysis Datasets

In addition to in-situ and remotely sensed datasets, reanalysis data sets were also analyzed. Reanalysis data is a combination of in-situ, remote sensing, and model data that are optimally combined using data assimilation techniques to provide a temporally and spatially consistent data set of environmental variables. For this analysis, the North American Land Data Assimilation System (NLDAS) data set was used. NLDAS is a reanalysis data set derived from the North American Regional Reanalysis (NARR) (Xia et al. 2012). The dataset has an hourly timestep and approximately a 12 km spatial resolution. The data used in the analysis includes precipitation, air temperature, shortwave radiation, longwave radiation, relative humidity, and windspeed in the north-south and east-west directions. NLDAS data was used from January 1, 2002 to December 31, 2021 and was temporally averaged to a daily timestep. The windspeed magnitude (WS_{mag}) was calculated from the north-south (WS_{NS}) and east-west (WS_{EW}) windspeed values with Equation 1.

$$WS_{mag} = \sqrt{WS_{EW}^2 + WS_{NS}^2} \quad [1]$$

Reanalysis data provides a useful source of environmental indicators that have continuous spatial converge that allows for widespread application.

2.3 Methods

The methods used in this analysis include calculating vegetation indices, statistical tests for trends and correlation and predictive model techniques including regression, clustering and regression trees. Each of these techniques are discussed in detail below.

2.3.1 Greenness Indices

Greenness or Vegetation indices detect the "greenness" of the land surface using different reflectances captured by satellites (Jiang et al. 2008). Since CyanoHABs are like vegetation in that they appear green, these different vegetation indexes were identified to explore their suitability as potential indicators. The color of green vegetation reflects red and near-infrared (NIR) radiation, so as chlorophyll becomes more dense, more visible light can be used by the algae to photosynthesize (Camps-Valls et al. 2021). The NDVI and EVI take advantage of this by using the red and NIR bands to create a measurement of this greenness (Camps-Valls et al. 2021). The main limitation of the NDVI it is nonlinear, and it becomes saturated over dense areas of vegetation, which the EVI tries to compensate for to some extent though it still experiences problems with saturation (Camps-Valls et al. 2021). The floating algae index (FAI) was also explored to see if it could provide more or better information than the NDVI and EVI because it was specifically developed as a method for detection of floating algae and has been shown to be potentially advantageous over the NDVI and EVI in detecting algae due to its ability to overcome atmospheric effects that plague the NDVI and EVI (Hu 2009) . The greenness indices used for analysis are averaged values across all open water pixels for each location.

The MODIS data set provides both the NDVI and EVI and is detailed in the MODIS Vegetation Index User's Guide (Kamel Didan, Munoz, and Huete 2015). Both the NDVI and EVI use the NIR and the red surface reflectance. For MODIS NIR uses band 2 (841-876 nm) and red uses band 1 (620-670 nm). NDVI only uses these two bands following Equation 2.

$$NDVI = \frac{NIR-red}{NIR+red} \quad [2]$$

The EVI developed out of the NDVI, building on the original equation to compensate for the deficiencies found in the NDVI (Equation 3). It additionally uses the blue band 3 (459-479 nm), a

gain factor for scaling (G), and three coefficients (C_1, C_2, and L). For the MODIS data set, G = 2.5, C_1 = 6, C_2 = 7.5, and L = 1. The added blue band helps correct for aerosol influences in the red band, and C_1 and C_2 are coefficients added to the aerosol resistance term. L is the canopy background adjustment for correcting the nonlinear, differential NIR and red radiant transfer through a canopy. Additionally, the EVI has a backup algorithm when bright targets lead to incorrectly high EVI values (Equation 4). The EVI2 removes the blue band, which becomes saturated with bright targets; thus, removing it from the equation is necessary to avoid an extremely high EVI. The coefficients are adjusted for this equation, with C_3=1 and L remaining the same. For both the NDVI and EVI, the bands are atmospheric corrected for Rayleigh scattering and ozone absorption.

$$EVI = G * \frac{NIR - red}{NIR + C_1 * red - C_2 * blue + L} \quad [3]$$

$$EVI2 = G * \frac{NIR - red}{NIR + C_3 * red + L} \quad [4]$$

The MODIS data set does not include calculated values for any other vegetation indexes. The FAI is calculated from MODIS reflectance data using bands for the NIR, red, and shortwave infrared (SWIR) bands averaged over the lake area. Like the NDVI and EVI, the FAI uses band 2 (NIR 841-876 nm) and band 1 (red 620-670 nm). The SWIR uses band 5 (1230-1250 nm). The equation developed by Hu uses a manually calculated reflectance that corrects for Rayleigh scattering; however, the MODIS surface reflectance used already corrects the bands for atmospheric conditions, including Rayleigh scattering, so this was unnecessary. For MODIS, the coefficients in this equation are λ_{NIR} = 859 nm, λ_{red} = 645 nm, and λ_{SWIR} = 1240 nm. Using these, the baseline reflectance for NIR band and calculated, and the difference between the NIR band and the baseline outputs the FAI (Equations 5 and 6).

$$FAI = R_{rc,NIR} - R'_{rc,NIR} \quad [5]$$

$$R'_{rc,NIR} = R_{rc,red} + \left(R'_{rc,SWIR} - R'_{rc,red}\right) * \left(\frac{\lambda_{NIR} - \lambda_{red}}{\lambda_{SWIR} - \lambda_{red}}\right) \quad [6]$$

For simplicity and ease of calculation, only NDVI was calculated for the LandSat data set. As the reflectance bands varied from satellite to satellite, there are slight differences in the bands used to calculate the NDVI. Table 2 shows the band and wavelength for each of the bands used for calculating the NDVI for different LandSat satellites.

Table 2. LandSat bands for NDVI calculation

LandSat Satellite	Red	NIR
LandSat 4 and 5	Band 3 (0.63-0.69)	Band 4 (0.76-0.90)
LandSat 7	Band 3 (0.63-0.69)	Band 4 (0.77-0.90)
LandSat 8 and 9	Band 4 (0.64-0.67)	Band 5 (0.85-0.88)

2.3.2 Trend Tests and Correlations

The Mann Kendall Trend test (MK test) is a nonparametric test for presence of a trend in a timeseries (Mann 1945). Using the pyMannKendall Python package, two trend tests were performed on the vegetation indexes, air temperatures, and surface temperatures at all reservoirs. The seasonal MK test and the original MK test, both at a 95% significance level, were used. The original test does not consider seasonal effects, while the seasonal test can be used with seasonal time series to identify seasonal trends (Mahmud n.d.). Trends were analyzed for daytime surface temperature, nighttime surface temperature, daily average surface temperature, air temperature, NDVI, EVI, FAI, and windspeed for each location. For the original test, seasonal effects were removed by looking at trends for each month separately. That is, trends between January of each year were analyzed with the test, and similarly for all other months. Similarly, the months of April through October were separated and analyzed as a group. Using these months, the effects of early

spring, late fall, and winter were removed since the temperatures during those times are typically lower.

In addition to trends, the data sets were also analyzed for correlation among variables. The correlation was calculated between the vegetation indexes and the NLDAS forcing data at significance levels of 95% and 99% using the Pearson correlation to identify linear relationships between these datasets exist. The correlation between the NDVI, EVI and FAI were also analyzed to explore the relationship between the different indexes.

For Cheney Reservoir, additional correlation calculations were done. Using the cyanobacteria cell count, the correlation between the cell count and the EVI, land surface temperature, water temperature, air temperature, windspeed (USGS, NOAA, and NLDAS) were analyzed. Additionally, the temperature and windspeed were explored further by calculating the correlation between different rolling averages, rolling maximums, and rolling minimums and the cell count to identify the optimal time scale for assessing conditions leading up to blooms that may play a key part in the prediction of CyanoHABs. Rolling averages of the independent variables were tested to identify if conditions leading up to the date of cyanobacteria measurement played a part in predictability. To do this, for each NLDAS and MODIS dataset, correlations between the cyanobacteria were calculated for varying numbers of averaging periods for the independent variable, up to 300 days.

Due to the uncertainty in the sediment core data, traditional correlation methods may not provide a representative assessment of the relationship. Therefore, a new methodology of assessing the relationship between predictions and the sediment core data was developed. To account for the large amount of uncertainty in the dating of the sediment core data, an iterative approach that relies on assessing the relationship through the median correlation and median significance level was

developed. For each iteration, the year associated with the pigment data was varied by adding a random integer between -2 and 2 to the year. If the newly generated year overlapped with the measurement before it, 1 year was added or subtracted to ensure they remained in the same order. After perturbing the sediment core data in this way, the correlation and significance between the sediment core data and the analyzed variable was calculated and stored. This process was iterated 2000 times for each relationship. The idea behind this process is that if there is an underlying relationship between the sediment core data and the analyzed variable, this will be captured by the median correlation across all iterations. When applying this technique with the sediment core pigment data, the correlation was found between three different pigments with the vegetation indices and NLDAS data. The vegetation indices and NLDAS datasets were averaged to a single yearly value to be temporally consistent with the sediment core data and only included data from May through September to incorporate the months with the highest potential for algae blooms when analyzing the relationship between these datasets and the pigments.

2.3.3 Nonlinear Regressions

Since Cheney Reservoir has estimated cyanobacterial cell counts available, more in-depth analysis with regressions was performed to better examine the relationship with satellite data. For all regressions, the cyanobacteria cell count is taken as the dependent variable, either at a daily or a three-day rolling average resolution. Regressions were completed using the Python module statsmodels using ordinary least squares (OLS) (Seabold and Perktold 2010). They were done for the entire year and for April through October, again accounting for the warmer months in the year.

Initially, the main approach was a linear regression, both with and without a constant. Single variable regressions with EVI, land surface temperature, water temperature, air temperature, windspeed (USGS, NOAA, and NLDAS). The regression module outputs the y-intercept (b) and

a single coefficient (x) (Equation 7). Multivariable linear regressions were performed with different combinations of these variables as well, with up to two variables (x_1 and x_2). In this case, the output includes a coefficient for each independent variable (m_1 and m_2) (Equation 8).

$$y = mx + b \quad [7]$$

$$y = m_1 x_1 + m_2 x_2 + b \quad [8]$$

Nonlinear regressions were also explored. The same statsmodel module was used, but the cyanobacteria cell count (y) was transformed by taking the natural log of the variable (y') before running the regressions which results in an exponential relationship (Equation 9).

$$y = e^{mx+b} \quad [9]$$

Multiple variables were incorporated into the nonlinear regressions, though not as a multiple variable regression. Colormaps with surface temperature, cyanobacteria cells, and windspeed were used to visually identify possible thresholds in either the temperature or windspeed that may allow for creating two separate regression equations. In both the linear and nonlinear regressions, equations with and without coefficients (b) with no coefficients were used. However, the regression equation's correlation output by statsmodel for these was often artificially inflated to values near 1 or were generally inaccurate when done without a coefficient. Therefore, the R^2 value for regressions with no coefficient are not used in evaluation.

The regression models were validated using the sediment core data and the same iterative technique as with correlations to account for uncertainty in the pigment data. This provides a means for quantifying the transferability of the nonlinear regression models as the regression models were developed at Cheney and then applied at the sediment core locations. This was done by calculating the annual average value from the regression model using only values calculated from April

through October. Then, the median correlation and median significance between the modeled fPC and pigment data was calculated over 2000 iterations.

2.3.4 Clustering

To further develop the regression models and find thresholds more specifically for possible regression equations, a clustering algorithm was used prior to performing a regression with the temperature. Windspeed, precipitation, shortwave radiation, EVI, and NDVI were used separately and in different combinations using K-means clustering with the Scipy Python library to find potential clusters in the data (Virtanen et al. 2020). To do this, data is first whitened, which normalizes the observations, using the scipy module for whitening, and then the K-means algorithm searches for clusters in the data by identifying centroids in a set number of clusters. The groups of data, or clusters, are separated by the algorithm trying to minimize the within-cluster sums of squares. The number of clusters can be varied, and in this case, sets of two or three clusters were used. Once centroids in the cluster are defined, datapoints are separated into one cluster based upon their distance from the centroid. After clusters are identified, a nonlinear regression was run on each separate cluster to produce multiple regression equations.

For the final model, a confidence interval was constructed around each cluster's regression line, and points outside of the confidence interval were removed. The confidence interval was constructed based on uncertainty in estimating the mean of the sample and slope regression parameter based on assuming a normal distribution of the regression errors. After removing points that fall out of the confidence level, new regressions were run for each cluster. Removing points outside of these intervals allows points that are likely representative of outliers at Cheney to be eliminated, making the model more transferrable to Marion since those outlying points may not well represent more general conditions. The final confidence level was selected by running the

regression at multiple confidence levels, then applying each at the sediment core location to test how this affected the correlation between the predicted values and pigments.

2.3.5 Regression Trees

As a final approach for creating a predictive model, regression trees were used with each of the datasets. Regression trees are a supervised machine learning method that creates a series of rules that form a predictive model. In regression trees, the decision node is created by selecting a separation decision (i.e. $X1 > 0.5$) based on the input variable and threshold that minimizes the variability in the two resulting data sets. This process is continued over and over again creating expanding branches that result in either another node or a leaf. A leaf is the end point of a regression tree and is meant to represent a series of data points that have similar properties based on the inputs. The predictive value of a leaf then becomes the average value of all the data points contained in the leaf. Thus, if a regression tree has more leaves, it will have more predictive resolution, but may come at the cost of over fitting the model. There are two key parameters that control the tree size and number of leaves created in the regression tree. The first parameter is the maximum depth. The depth of the tree is a measure of levels of decisions. Thus, a regression tree with a depth of 1 will only have one decision node that results in two leaves. Given the bifurcation nature of the decision tree, as the depth increases, the number of leaves increases. The other important parameter for the algorithm is the minimum sample size in a leaf. This parameter controls the expansion of new branches by only allowing the creating of new decision nodes if the resulting leaves have the minimum number of values. Thus, increasing the minimum sample size of leaf will reduce the number of branches. Regression trees are advantageous because they eliminate the need for clustering data and allows for more in-depth predictions than using sets of clustered data in nonlinear regressions. Their output is also simple and easy to understand, as the

decision tree can be plotted to show the decision at each node and followed until it arrives at the predicted value. It is also helpful because it can better manage outliers and classify those values in a way that clustering algorithms may not be able to capture. As with any machine learning algorithm, there is always the potential for overfitting the data. To help mitigate this, the two key parameters (maximum depth and minimum leaf sample size) of the algorithm were optimized by breaking up the data into 80% training and 20% validation. The maximum depth and minimum leaf sample size were chosen based on the ones that maximized both predictive variance and correlation within the validation data set. Estimating these parameter values based out of sample a data set should help reduce the chance of overfitting the model to the data. To evaluate the importance of predictors in the regression trees, the feature importance, an attribute output by the same Python module, is also calculated for each feature in the regression tree. This attribute tells which predictor inputs are most important to prediction, with higher values indicating more importance. It is calculated as the decreased in node impurity weighted by the probability of reaching a particular node. To calculate the node probability, the number of samples in a node is divided by the total number of samples.

Chapter 3: Results

3.1 Environmental Characteristics of Kansas Reservoirs

Before developing predictive models of CyanoHABs, the different environmental variables are first assessed for trends and relationships through exploratory analysis. This exploratory analysis is crucial for identifying key variables that will drive the predictive models.

3.1.1 Trends in Kansas Reservoirs

The entire MODIS temperature dataset, from July 2002 to July 2021, was first analyzed with the Mann Kendall trend tests. On a monthly basis, significant daily average daytime LST

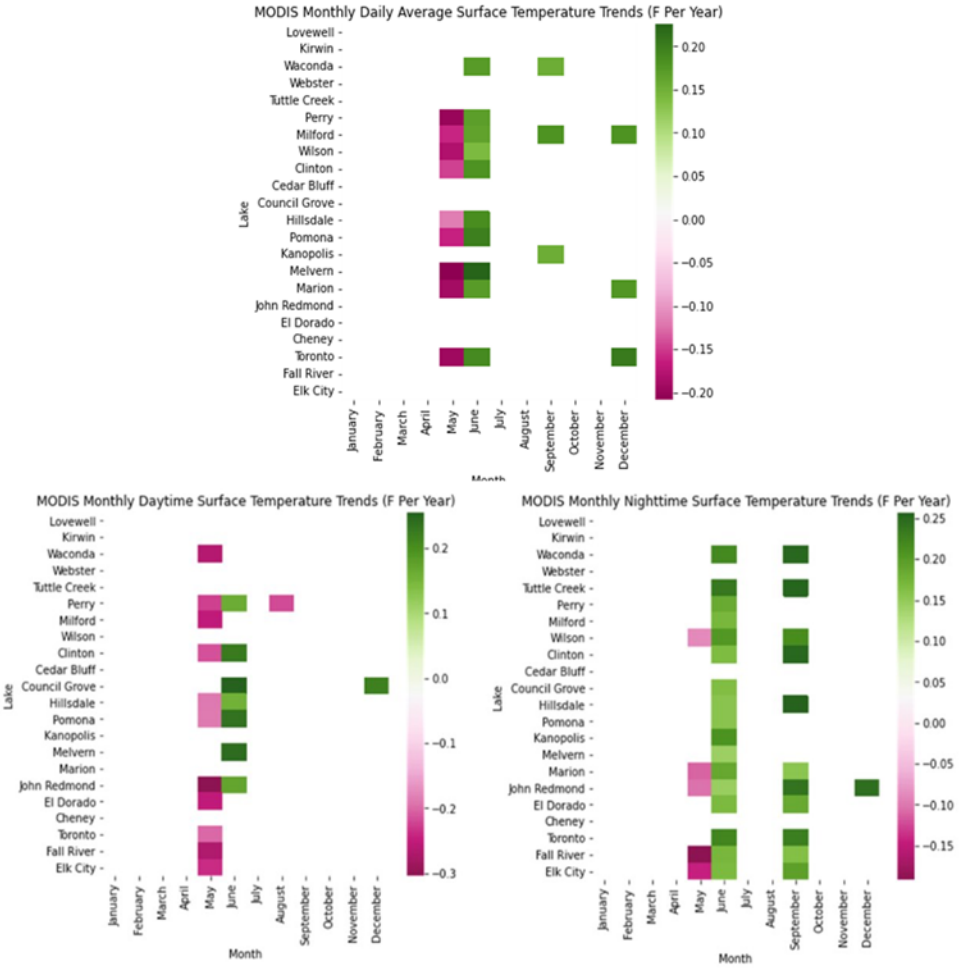

Figure 4. Monthly MK trend test results for MODIS LST. Larger positive values indicate a positive trend and larger negative values indicate negative trends. Only values that are statistically significant at a 95% confidence level are shown.

trends from MODIS were only seen in May, June, August, and December, with consistently decreasing temperatures across most locations in May and increasing temperatures in June across several locations (Figure 4). The decreasing water surface temperature trends in May do not have a strong explanation. Similar tests were performed on NLDAS air temperature, precipitation, windspeed, and shortwave radiation to understand why it is occurring. None of the locations had decreasing air temperatures and increasing windspeeds, and only three locations had increasing precipitation (Webster, Kanopolis, and Milford) and decreasing shortwave radiation (Lovewell, Perry, and Milford). Air temperatures decreasing would cool the surface, and windspeeds increasing would indicate more mixing, bringing up cooler waters from the deeper parts of the lake. Increasing precipitation would add cold water to the reservoir and decreasing shortwave radiation would mean less sunlight reached the surface. Since none of these things are occurring, there is little explanation for the decreasing trend in water surface temperature. Monthly nighttime

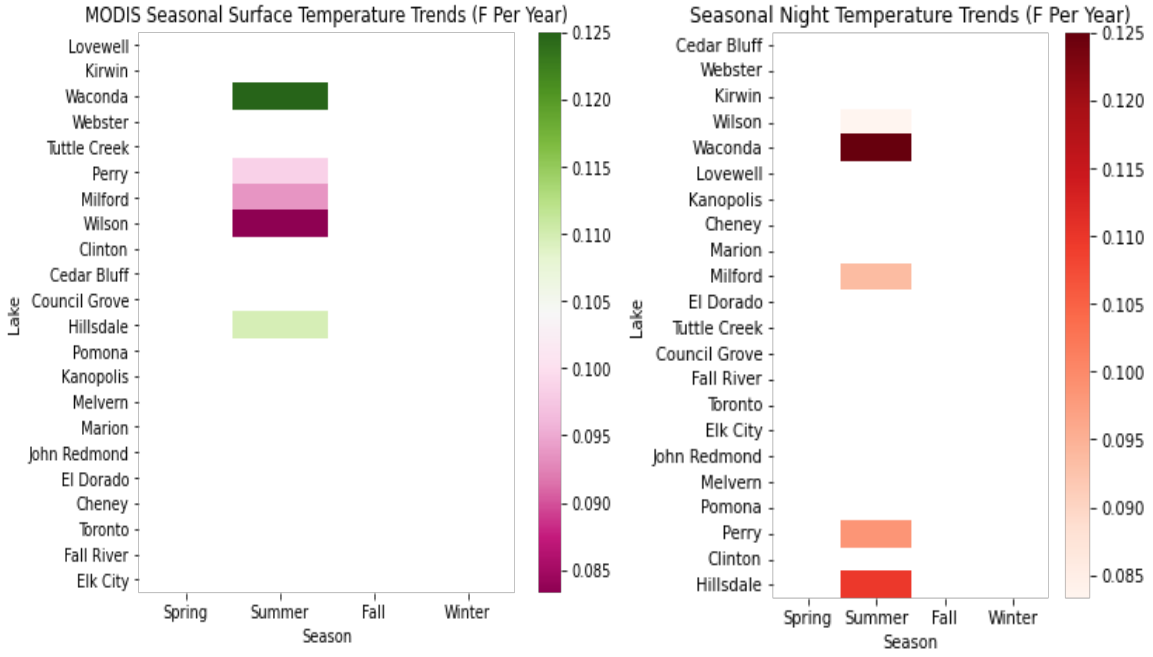

Figure 5. Seasonal MK trend test results for MODIS LST. Larger positive values indicate a positive trend and larger negative values indicate negative trends. Only values that are statistically significant at a 95% confidence level are shown.

trends had a similar pattern in May and June, but additional consistently increasing trends in September. Since HABs thrive in warmer temperatures, these increasing temperatures could contribute to increasing blooms earlier in the spring and extend blooms later in the summer. With nighttime temperatures increasing in the summer months, this may provide even more ideal conditions for bloom growth throughout the summer by creating more consistently warm temperatures for bloom growth. On a seasonal basis, average MODIS LST increased at only five locations (Figure 5). During the day, these seasonal trends did not appear, but during the night five locations saw warming temperatures during the summer. This again illustrates that the year-to-year conditions are becoming more ideal for bloom growth during their most prominent growing season for some locations. These trends, and the following seasonal trends, were analyzed considering north-south and east-west positions of the reservoirs, but no clear location-based patterns emerged.

The entire MODIS dataset was again used for seasonal trends, with seasons defined in three-month increments, starting with summer as June, July, and August. Seasonal trends for the remote sensing greenness indices using MODIS are shown in Figure 6. For FAI and EVI, several locations had increasing trends in the summer. Since these measurements are over open water, this would seem to indicate that the greenness in the water is increasing during the summer months, potentially showing a relationship with increased algae mass as it increases during the warmer season. Fewer consistent trends appeared in other seasons. For the FAI only, Webster and Kirwin reservoirs, both had decreasing trends in FAI in all seasons. The NDVI did not show a clear pattern of seasonal trends for multiple locations in the summer or fall. Four had increasing trends in the spring, and five reservoirs had increasing NDVI in the winter. In the spring, this could indicate

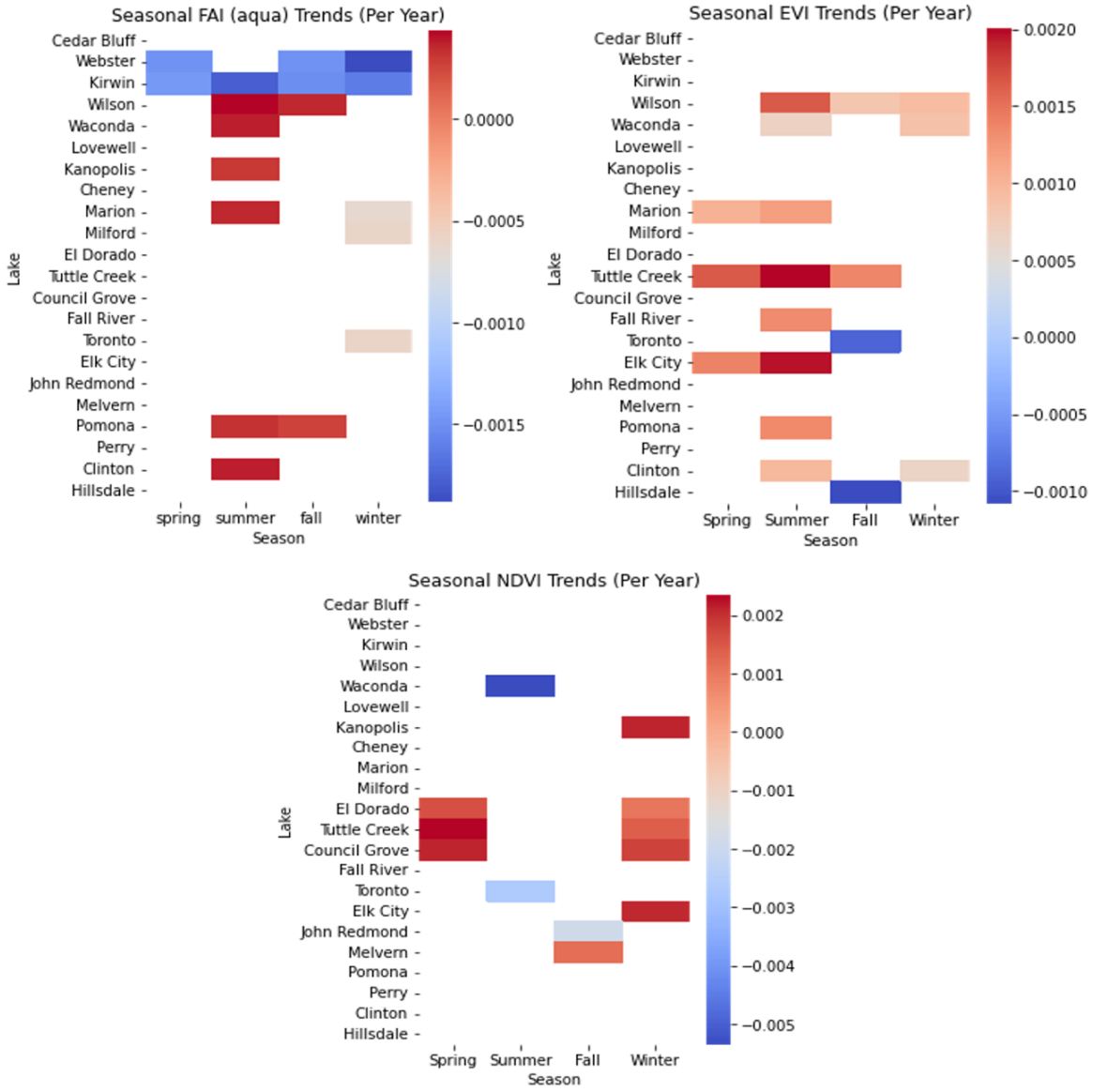

Figure 6. Seasonal MK trend test results for vegetation indexes with reservoirs ordered west (top) to east (bottom). Large positive values indicate a positive trend and larger negative values indicate negative trends. Only values that are statistically significant at a 95% confidence level are shown.

more algae in the water as the weather warms up, but the increasing NDVI in the winter has a less clear source.

3.1.2 Correlation between Variables at Kansas Reservoirs

The relationship between the different greenness measures was explored at all locations to test their consistency. For the MODIS Aqua satellite, NDVI did not correlate well with the FAI.

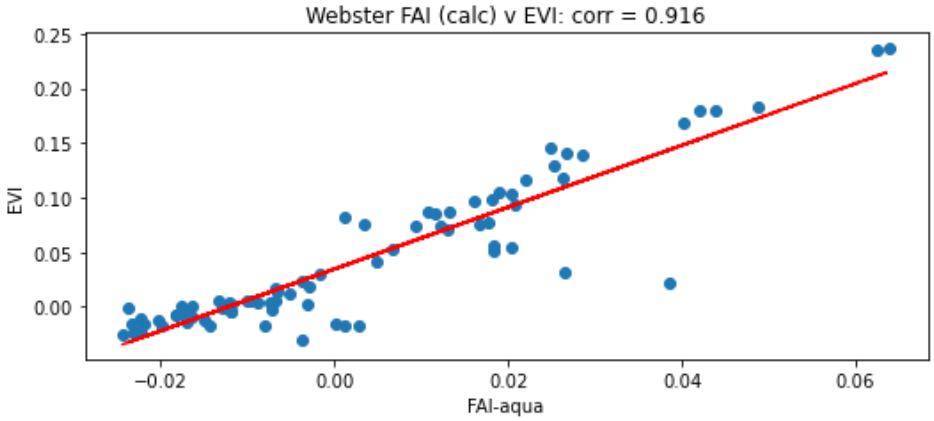

Figure 7. Correlation for EVI and FAI at Webster Reservoir.

EVI had a correlation with the FAI greater than 0.5 at all locations except Lovewell and Waconda. The highest correlation was 0.915 at Webster (Figure 7). Similarly, for the Terra satellite the EVI was well correlated with the FAI at all locations except Lovewell and Waconda, and the NDVI did not have a strong relationship with the FAI. From this, the vegetation indices do seem to relay some of the same information, but differences in how and which surface reflectance bands are used in their calculation is likely leading to variation in the greenness values.

Next the relationship between the remotely sensed greenness measures are compared to other environmental variables from the NLDAS reanalysis. The vegetation indexes tended to be positively correlated with the NLDAS datasets (Figure 8). Temperature, shortwave radiation, longwave radiation, and relative humidity had the highest correlation for all indexes at most locations. Since warmer temperatures are occurring when we expect the greenness to be the highest, the positive correlation between these variables indicates a relationship between temperature and greenness that is consistent with expected algae growth. The EVI at all locations had a greater correlation than the others with almost every NLDAS variable, though NDVI was not much behind. The FAI did not have as strong of a relationship with as many variables at as many locations. This indicates that the NDVI and EVI are picking up on a similar source of variability as the NLDAS variables while the FAI has a source of variability that is more unique

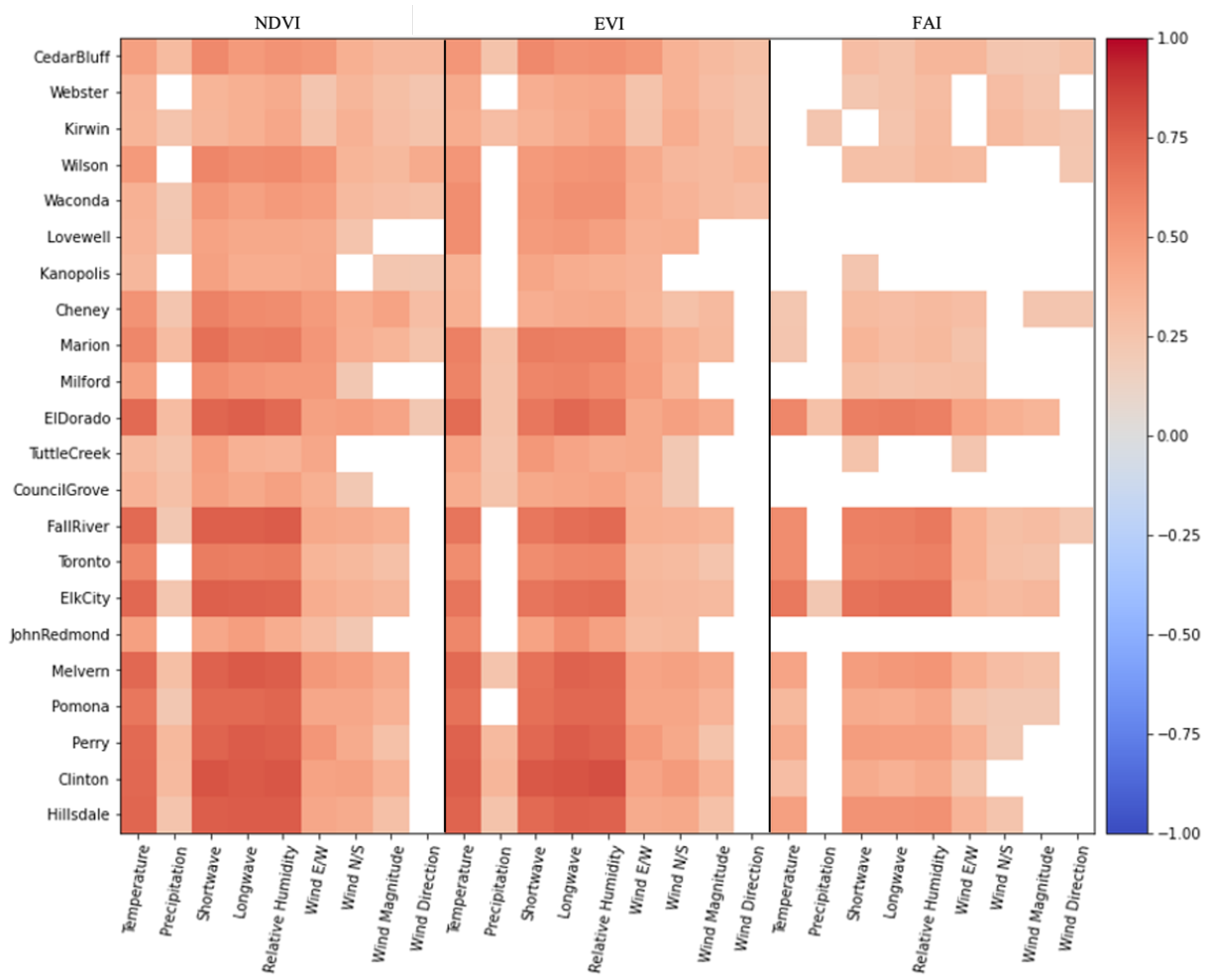

Figure 8. Correlations between EVI/NDVI/FAI and NLDAS datasets. Only values with statistically significant at a 95% confidence level are shown.

compared to the NLDAS variables. Whether or not these signals are linked to cyanobacteria or some other source of greenness will be explored in the next section.

3.2 Identifying Potential Predictors

To identify specific predictors at Cheney Reservoirs, correlations between cyanobacteria data and the different satellite parameters were tested. Additionally, more in-depth relationships between windspeed and temperature were explored to determine if more than one relationship was necessary to model the cell counts. This testing is important to determining what initial inputs into later regressions will be most effective in building a model.

3.2.1 Correlation with cell count data

For Cheney Reservoir, the existence of cyanobacteria data allowed for more in-depth testing of relationships with the different environmental variables from satellite and reanalysis. For this analysis, the satellite variables and the NLDAS variables are compared to the daily average cell count as measured by the high frequency sensor from October 1, 2012 to March 12, 2015.

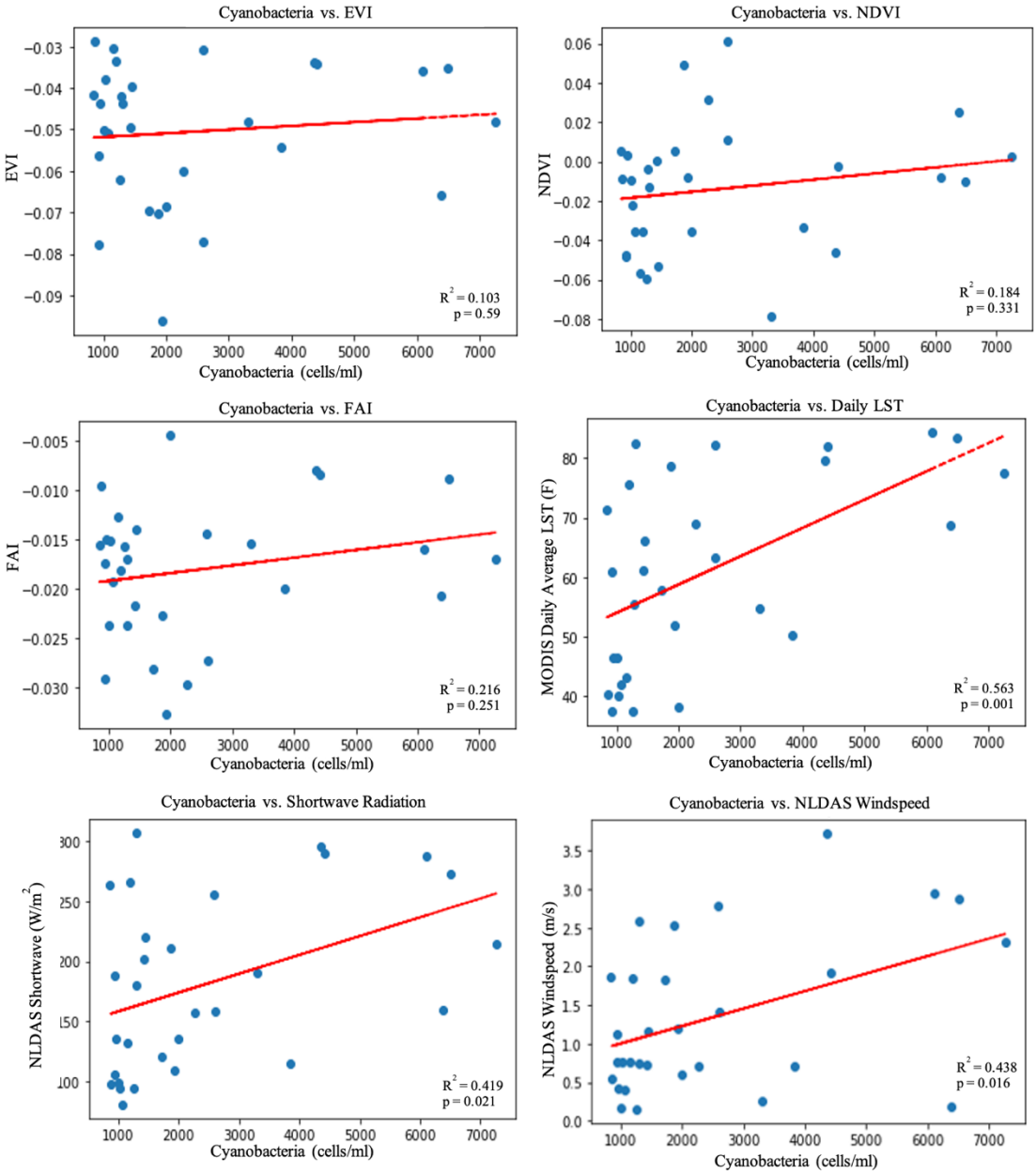

Figure 9. Correlations with Cheney Reservoir cyanobacteria cells counts with daily average MODIS LST, MODIS vegetation indexes, FAI, and NLDAS shortwave radiation. Only trends with < 0.10 correlation are shown, regardless of significance level (bottom right corner of plots).

Here, the LST and shortwave radiation had the highest correlation and were significant at a 95% confidence level. Both of these variables showed a strong relationships with other NLDAS data and the greenness measures (Figure 9). Despite this, the greenness measures were not significantly correlated to the cyanobacteria data, though of the three, FAI had the highest correlation. The lack of a relationship also may be due to the cyanobacteria being point measurements in the reservoir, while the greenness indices are spatially averaged across the water surface. Interestingly, the EVI shows less of a relationship to cyanobacteria at Cheney as compared to the FAI and NDVI, despite prior relationships appearing to show EVI having a stronger relationship with other environmental variables. As a whole, the greenness measures were low and not significant. This indicates that the greenness measures may not be good indicators of blooms, at least not without additional environmental drivers such as surface temperature, shortwave radiation and windspeed. In addition to the NLDAS data, wind speeds from in-situ measures were also explored. Of the windspeeds available for this location, the NLDAS windspeed had the highest correlations, and neither of the other windspeeds showed significant relationships with the cyanobacteria. However, the relationship between the NLDAS windspeed and cyanobacteria is not as expected as it indicates that as windspeed increases the cyanobacteria also increases. Increased wind speed increases mixing in the lake which should act to break up algae blooms. This likely indicates that the wind speed is actually capturing something different. It may be capturing the broken-up blooms as greenness spreads across the water surface, which could leader to higher overall greenness values for the reservoirs as more open water pixels become greener from the spread of algae. In contrast, the relationships found with temperature and shortwave radiation are consistent with the known relationship between cyanobacteria and temperature, as well as the availability of sunlight.

3.2.2 Temporal Scale of Predictors

The above analysis was based on comparing the different environmental variables and cyanobacteria cell count at the daily time scale; however, other averaging periods may help reduce the noise and be a better predictor. In this section, the optimal time scale for predictors is explored. When testing different averaging periods, a wide range of optimal averaging periods were

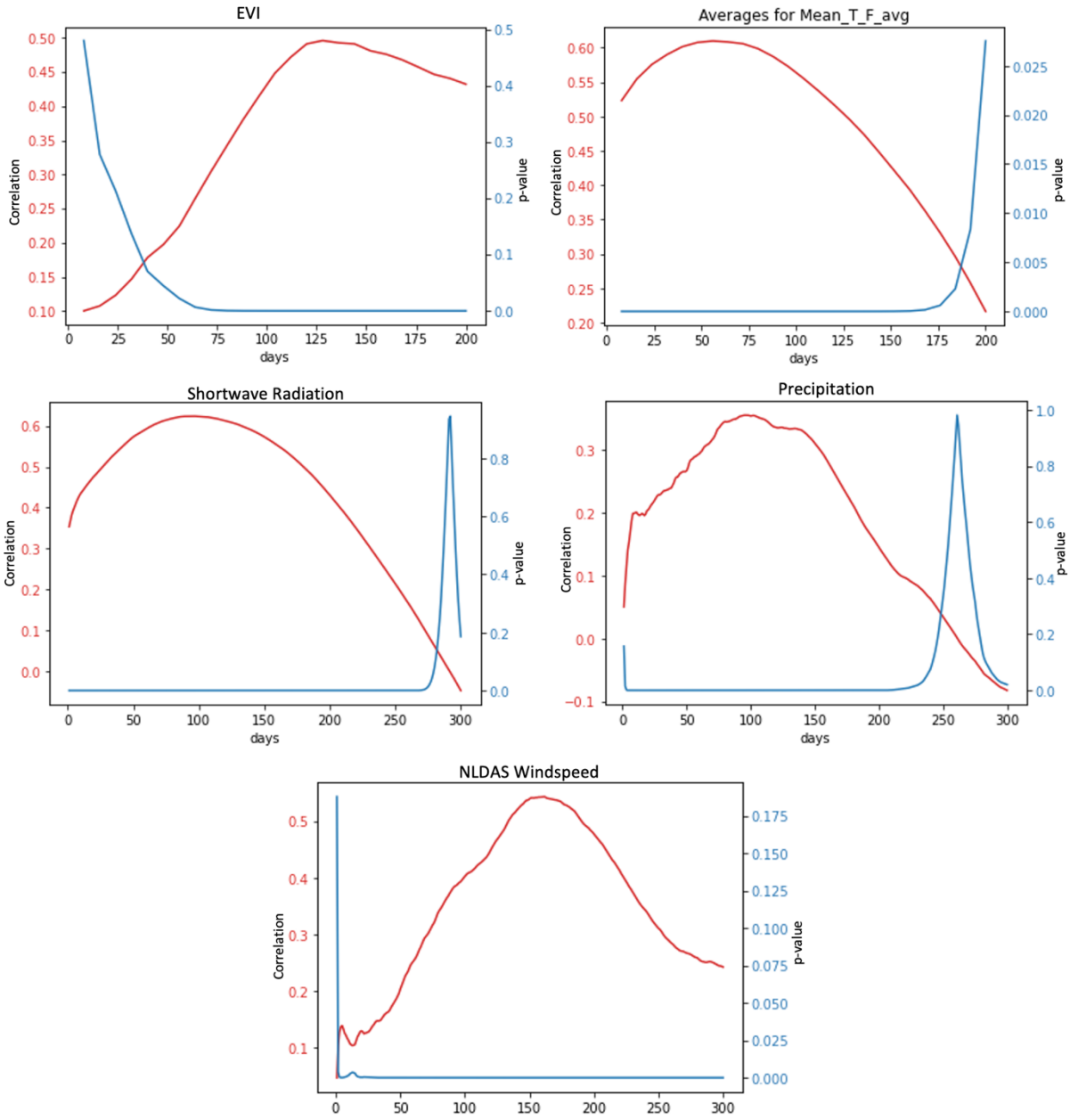

Figure 10. Correlation (red) and p-values (blue) for averaging periods for potential indicators. The peak of the correlation is taken as the optimal averaging period.

considered for the different inputs and the correlation between the variable at different averaging periods and the cell count are analyzed (Figure 10). The EVI had a maximum correlation at 128 days. The daily average LST and NLDAS temperature peaked at 56 days, while both the separate day and night temperatures peaked at slightly later at 64 days. The optimal timescale for shortwave was 96 days, relative humidity was 63 days, and precipitation was 97 days. The windspeed magnitude was 162 days, which was the longest of all the tested datasets. For comparison, the USGS water temperature had the highest correlation at a 45-day averaging period. The USGS windspeed had by far the longest averaging period, at 291 days at a correlation of 0.4; however, it also had a maximum negative of around -0.4, where the averaging period was 42 days. For all of the independent variables, the maximum averaging period was shortened to 32 days for consideration in the prediction models, despite potentially having a longer optimal averaging period. Doing so ensured that entire seasonal or near-yearly averages were not used, because this may not provide useful information (low resolution) about the conditions that lead up to blooms. While the longer averaging periods may not provide useful information, it shows the seasonal relationship that exists with blooms at Cheney Reservoir. Specifically, the longer averaging periods likely indicate more seasonal drivers of blooms than shorter-term relationships in the available predictors. The conditions during these longer periods can set up for bloom growth, as longer periods of lower winds would allow for more lake stratification, longer periods of higher temperature would warm the water surface more, and the longer-term effects on nutrient loading from precipitation events could take effect.

3.2.3 Threshold Relationships

Thus far, one of the interesting results about potential indicators that needs further exploration is the lack of relationship between cyanobacteria cell counts and wind speed. Since

windspeed did not have a strong relationship on its own with the data, potential threshold relationships are explored by considering the relationship between surface temperature, wind speed and cyanobacteria cell counts. This was done by separating the dataset out by temperature thresholds to determine if there was a more complex relationship between lower windspeeds and increase cell counts of cyanobacteria (Figure 11). No apparently relationship existed in this, but it could be too complex for this method of analysis or not be playing a large role in bloom growth at this location. There does appear to be some relationship between higher windspeeds and lower values of cyanobacteria, and with the known relationship between low-winds and blooms, this is logical; however, the relationship was not clear at higher temperatures.

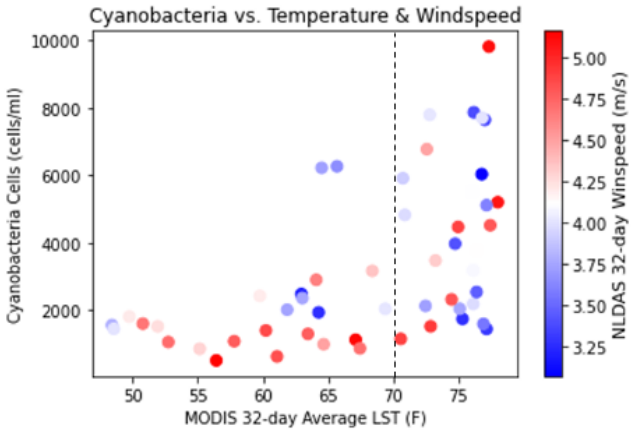

Figure 11. Colormapping of 32-day NLDAS windspeed over cyanobacteria cell count versus MODIS 32-day average LST with line denoting where relationship between temperature and windspeed begin to have a less clear relationship (70°F).

3.2.4 Sediment Core Data and Potential Predictors

The correlations between pigments and other variables varied greatly between different reservoir and sediment core sampling locations. At Marion Site 1, the highest correlation was found with the EVI, where R^2 varied between 0.3 and 0.48 for the different pigments, the highest being zeaxantin-lutein. The NDVI did not have a strong relationship (Figure 12). The temperature had a similar range of correlations, again having the highest with the same pigment as EVI. At Marion Site 2, the temperature had the highest correlation, though this time with canthaxthin. Milford Site 1 also had its strongest relationship with temperature and wind direction. At Milford

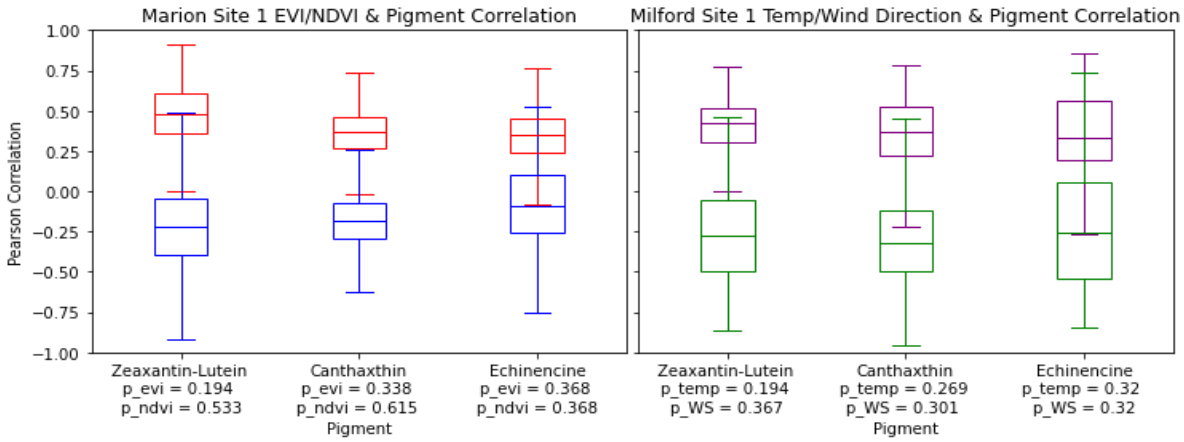

Figure 12. Box and whisker plot with median correlations with sediment core pigments at Marion (left: red = EVI, blue = NDVI; right) and Milford (right: purple = temperature, wind direction = green). Median p-value is included for each tested variable and pigment.

Site 2, the temperature had a high correlation and had the strongest relationship with the same pigment as Marion site 1. At Kanopolis, Perry, and Webster, there were no clear relationships with any variable tested. This indicates that different variables may be better predictors at one site, but not at another. However, there is some level of consistency between what is seen at Cheney and at Marion since the regression equations are able to be applied with improved Pearson correlations with this final regression equation. Furthermore, Cheney and Marion are also close geographically and have a similar orientation, which may result in a similar relationship between variables like wind and shortwave radiation. As such, the analysis on the prediction models will focus on fitting the model using the Cheney data and then validating it at Marion.

3.3 Prediction Models

Using the potential drivers identified above, different prediction models were created to test the performance of satellite data in predictive models. Linear regression, nonlinear regression paired with clustering, and regression trees were used as different approaches in this process.

3.3.1 Regression Models

The initial relationships established with temperature, shortwave radiation, windspeed, and NDVI provided key information about where to begin with regression models, and how different averaging periods could be made useful for them. Using single variable linear regression models with MODIS daytime average temperature (not a rolling average) showed the highest R^2 at 0.247. Other single variable and multiple variable linear regressions were tested, though these were not continued as initial nonlinear regressions showed higher correlations. The poor performance of the linear models indicates that there is likely more of a non-linear relationship between the algae data and the environmental variables that quantify bloom conditions. Nonlinear regressions were initially tested without constants. In analyzing these relationships, it appeared that two potential fits may exist: a lower area where the was a large amount of datapoints and an upper area that was somewhat separated. To attempt to find a threshold relationship and separate out two regressions lines from these relationships, thresholds were used to determine if specific wind conditions were ideal, and a threshold of 4.4 m/s resulted in the highest correlation (Figure 13). This did slightly improve from prior models and indicates that at least two regression lines may be useful in capturing the relationship between algae and surface temperature.

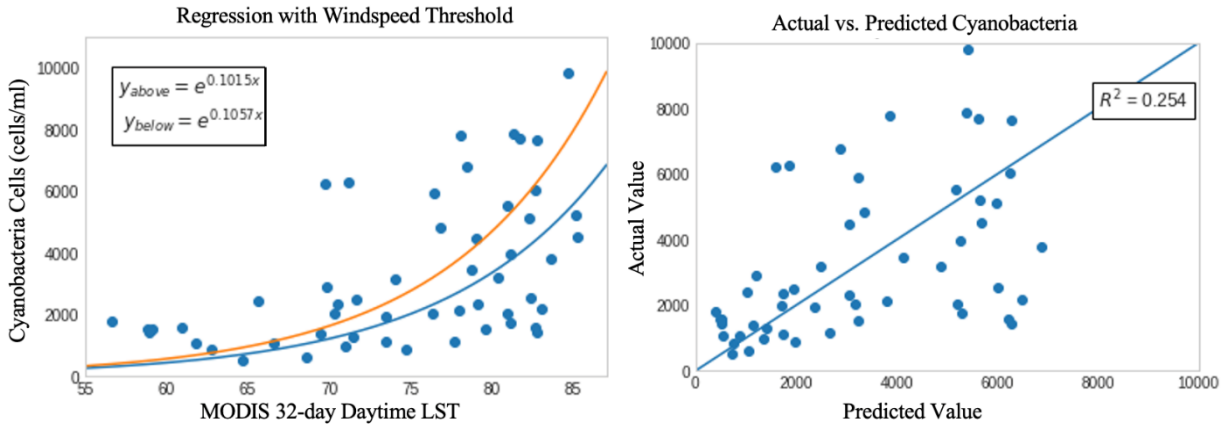

Figure 13. Nonlinear regression with data separated by a windspeed threshold of 4.4 m/s. No constant used in the regression equation, leading to artificially inflated regression correlations.

3.3.2 Regression Models with Clustering

Learning from initial regression models, a more robust way to subset data was needed to improve the regression models. Clustering allowed a better way to do this and was tested on all models with the different combinations of satellite data as well as the NLDAS forcing data to fully evaluate which predictors provide the greatest predictability.

NLDAS Model: The NLDAS regression, which consisted of air temperature, shortwave radiation, wind speed and precipitation had the largest dataset for the given period. This is due to the fact that the NLDAS data is available at a daily time step, while MODIS is only available every 8 days, though NLDAS may not be as representative of the water surface since it is gridded data.

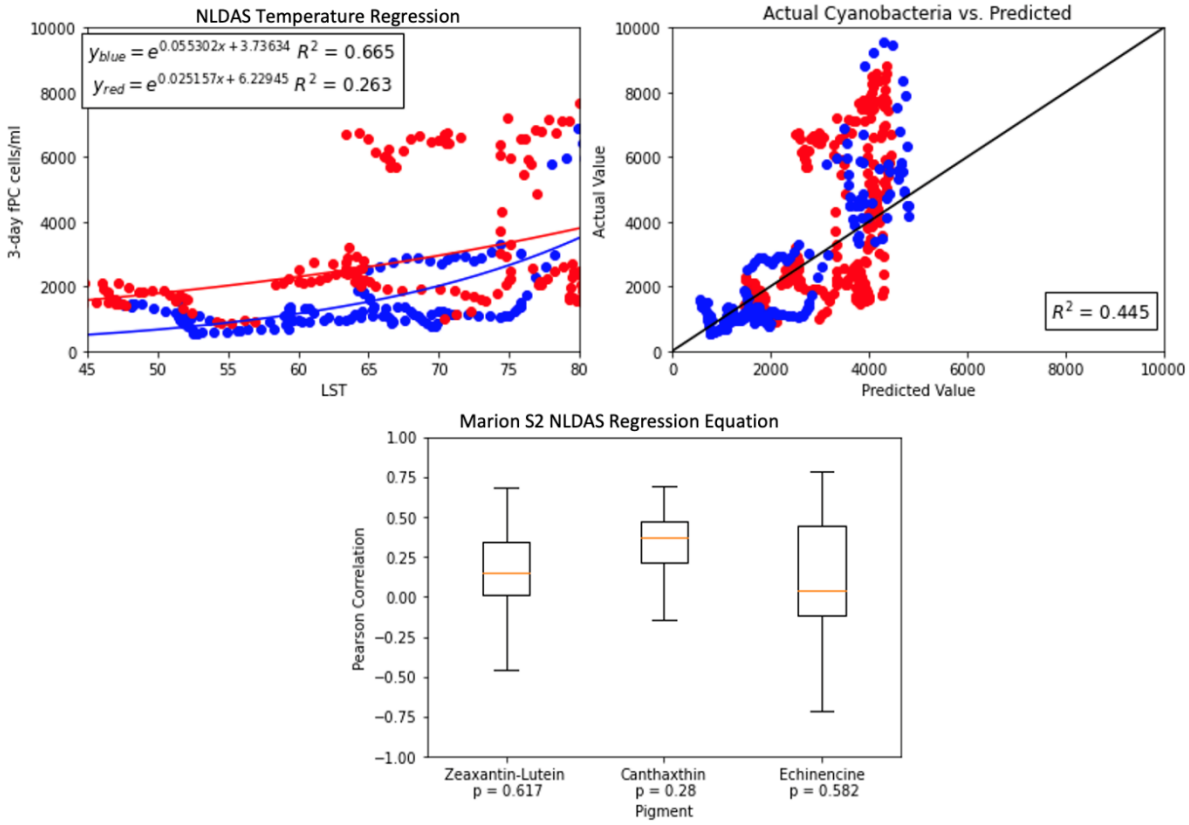

Figure 14. NLDAS final regression model with Cheney cyanobacteria. Two separate clusters with inputs of NLDAS 32-day averages of precipitation, windspeed, and shortwave radiation are included, and the regression is modeled using 3-day cyanobacteria data as the independent variable and NLDAS 32-day temperature as the dependent variable. The box and whisker plot shows application to the Marion pigment data with median correlation and median p-values listed.

The averaging periods of 32-days for the temperature, shortwave radiation, precipitation, and windspeed produced the best actual versus predicted relationship and were used for the in the final NLDAS model (Figure 14). The 3-day average of the cyanobacteria data was used with this as well, capturing the days leading up to the blooms. Two somewhat clear relationships emerged, with a cluster of primarily lower cyanobacteria values and one of higher values. When the regression equation was applied to the sediment core data at Marion, it was the highest for canthaxthin (R^2=0.312, p =0.28). Despite being the most temporally complete, the NLDAS data did not provide strong transferability to the sediment core data.

MODIS Model: Since MODIS was the initial satellite dataset for testing these relationships, it lent itself to more in-depth clustering tests in preparation for application with other datasets. The first attempt at using the clustering algorithm to separate data used one variable at a time, using the various NLDAS inputs available, then adding in second variables, all with 32-day rolling averages. This resulted in a two-variable cluster with wind and precipitation (Figure 15). The actual vs predicted correlation increased due to this addition of clustering, showing that this method improved upon the prior nonlinear relationships without it but still with room for improvement. With clear room for further improvement, constants were added to the regression

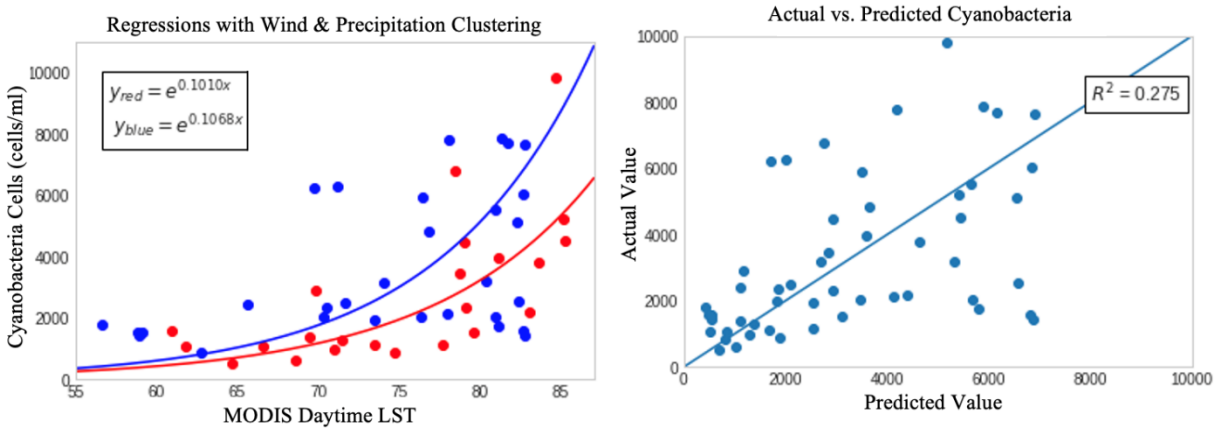

Figure 15. Nonlinear regressions with two clusters of NLDAS 32-day windspeed and precipitation. Daily cyanobacteria data is the dependent variable and MODIS 32-day LST is the independent variable. The regression equation does not include a constant, leading to artificially high regression correlations.

equations. The addition of constants not only allowed for the use of the regression correlation in conjunction with the actual versus predicted correlation but also led to immediate improvements in the model. This addition, plus the addition of shortwave radiation to the clustering algorithm increased the actual versus predicted correlation to 0.375, the largest increase up to that point with the model (Figure 16). The last variable altered was the cyanobacteria values. Rolling averages for 3-, 5-, and 7- day periods resulted in actual versus predicted correlations of 0.398 (Figure 16), 0.387, and 0.373, respectively. While this made some improvement, it is likely capturing the conditions of bloom growth more fully and allows for better prediction. The greenness indices were tested in the clusters as well. The EVI was tested at a 32-day averaging period with this

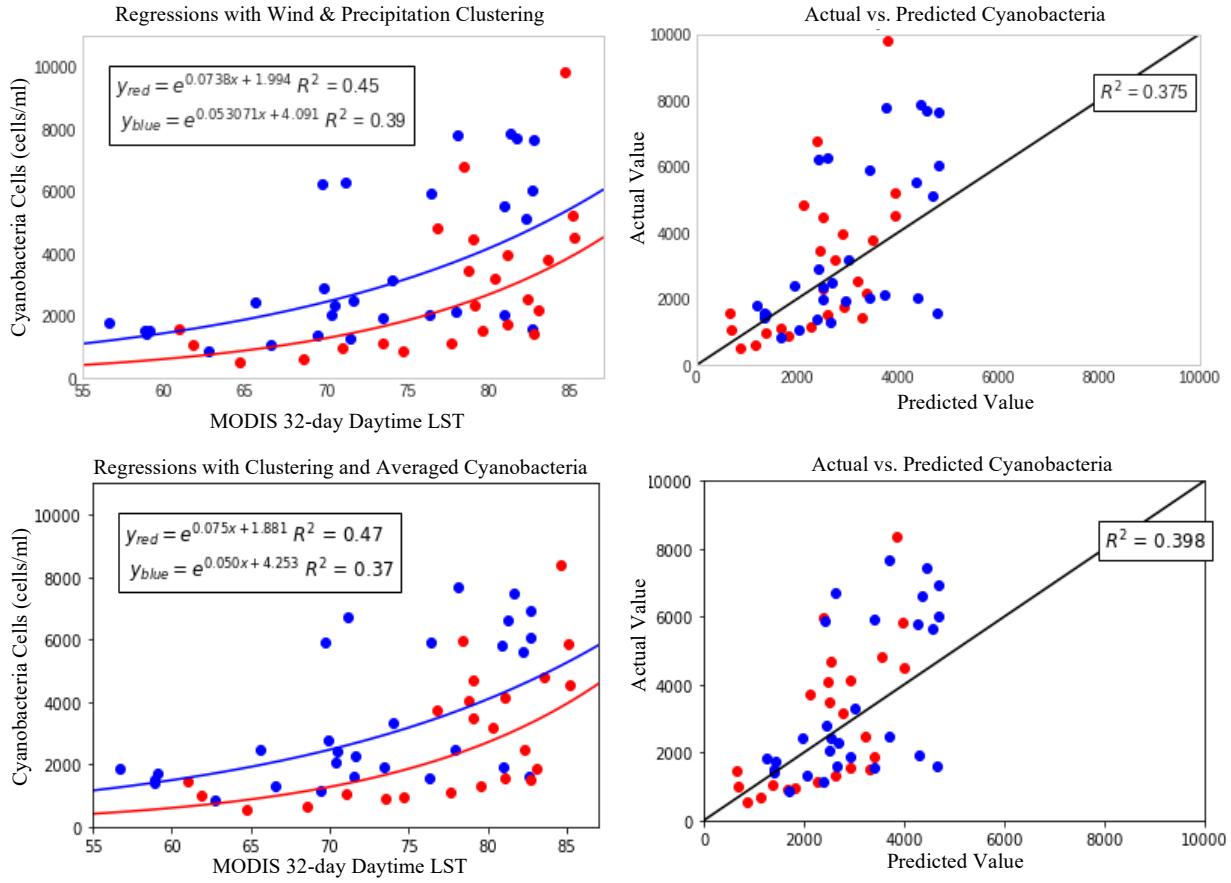

Figure 16. Nonlinear regressions with two clusters using NLDAS 32-day windspeed, NLDAS 32-day precipitation, and NLDAS 32-day shortwave radiation using daily cyanobacteria values as the dependent variable and MODIS 32-day LST as the independent variable (top) and the same model with 3-day rolling average cyanobacteria data as the dependent variable instead (bottom).

regression, and slightly decreased the actual versus predicted correlation to 0.384. It was also tested at other points in regressions, and results in slight decreases or no change in the correlation. The NDVI, also using a 32-day rolling average, resulted in a similar decrease in the correlation to 0.385. When using both the EVI and NDVI in regressions, the number of datapoints was reduced to about half due to the 16-day measurement increments by the satellite. From this, the greenness indices are likely either redundant with the temperature variable or not showing a strong relationship with cyanobacteria at this location, and, at least for Cheney Reservoir, do not strengthen the prediction framework.

To see if conditions leading up to blooms could be better represented by shorter averaging periods, using the windspeed, precipitation, and shortwave radiation, the averaging periods on the clustering variables were tested at 3-, 5-, 7-, and 14-day averaging periods when finding the strongest relationship. Looking at different combinations of averaging windows for each, the 5-day windspeed, 14-day shortwave radiation, and 32-day precipitation resulted in highest correlation (Figure 17). These shorter averaging periods for the shortwave radiation and temperature may be because the conditions in the two weeks and few days leading up to the bloom

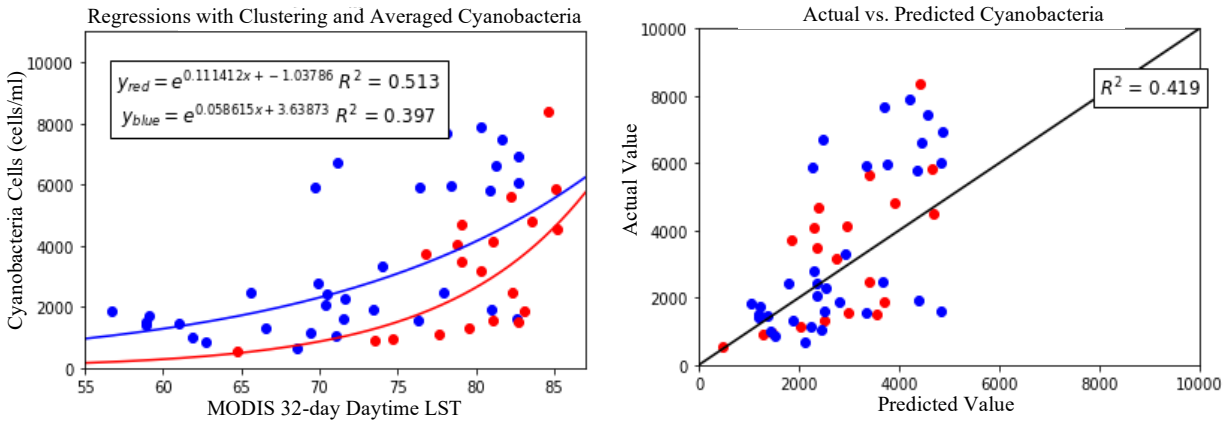

Figure 17. Nonlinear regressions with data separated by two clusters using NLDAS 5-day windspeed, NLDAS 32-day precipitation, and NLDAS 14-day shortwave radiation, with 3-day rolling average cyanobacteria cell count as the dependent variable and MODIS 32-day LST as the independent variable.

are more important for the impact of these variables, while the longer averaging period for the precipitation may better represent later increased nutrient loading from precipitation events. Additionally, the monitoring station in the reservoir is at a set level, meaning that it may be near the surface or deeper depending on the water level. If it is deeper, it may not capture the bloom as well since they are usually denser near the surface. Figure 18 shows the clustering of these variables. Windspeed and precipitation show the clearest distinct clusters, but less distinct clusters for the other inputs imply a more complex relationship. The relationships appearing in the clusters appear to be consistent with expected drivers of blooms: one cluster groups high precipitation with lower windspeeds, along with higher amounts of shortwave radiation. While the separation between the shortwave radiation and the other two variables in not as distinct, it shows these clusters were used with the remainder of regressions performed.

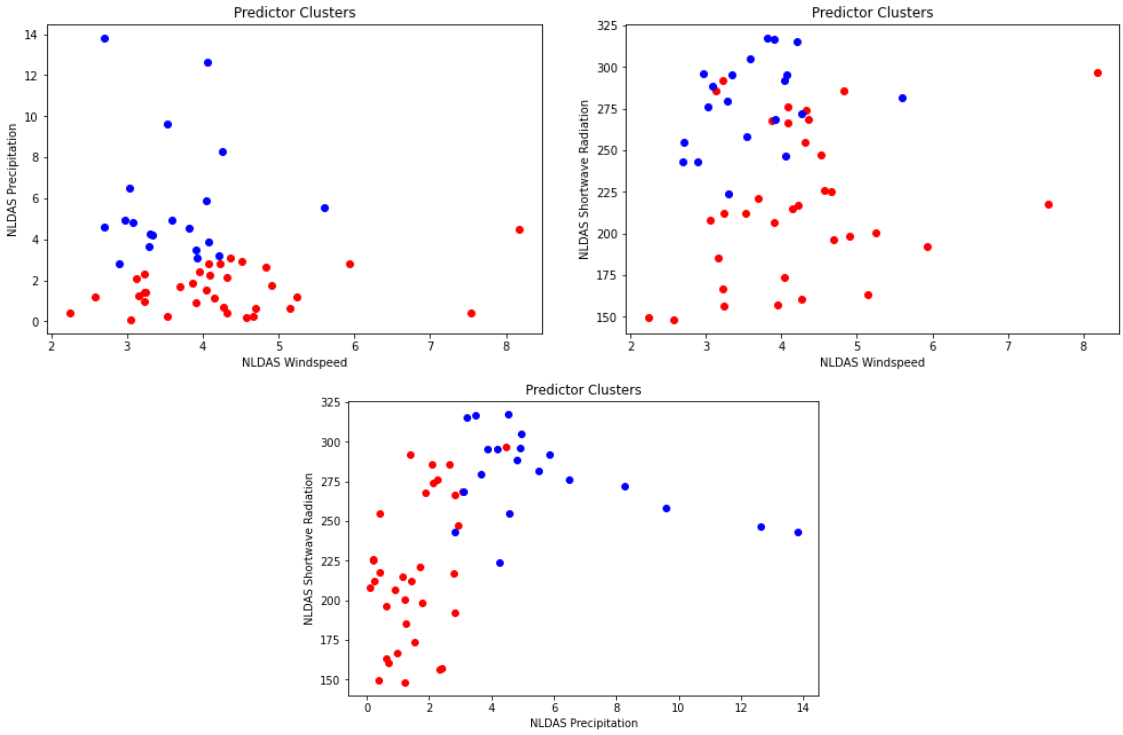

Figure 18. Predictor clusters for NLDAS 32-day precipitation, 14-day shortwave radiation, and 5-day windspeed each plotted against each other.

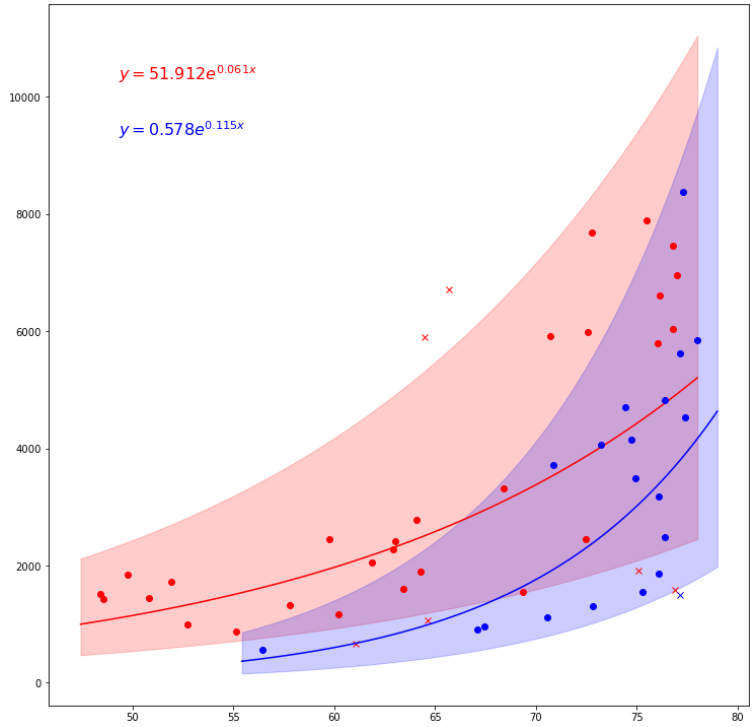

Figure 19. Regressions with 90% confidence intervals constructed around them. Points outside of confidence intervals are indicated as an x.

Using a confidence interval of 90% on the regression lines created by the final clusters, 7 total points were removed from the dataset and the same lines were replotted to see which points were removed (Figure 19). To see if these datapoints could be classified separately, a third cluster was added to the clustering algorithm and another regression line was added (Figure 20). The addition of a third cluster did not capture these points and the new third cluster only had 9 data points in it for the regression. After this did not address the issue, a third cluster was no longer used. To look deeper into the outlier points, they were removed from the dataset to see effects on the regression equation. Removing them improved the regression to a higher correlation than previously found, as well as the highest correlation so far for the actual versus predicted values (Figure 21). The removal of

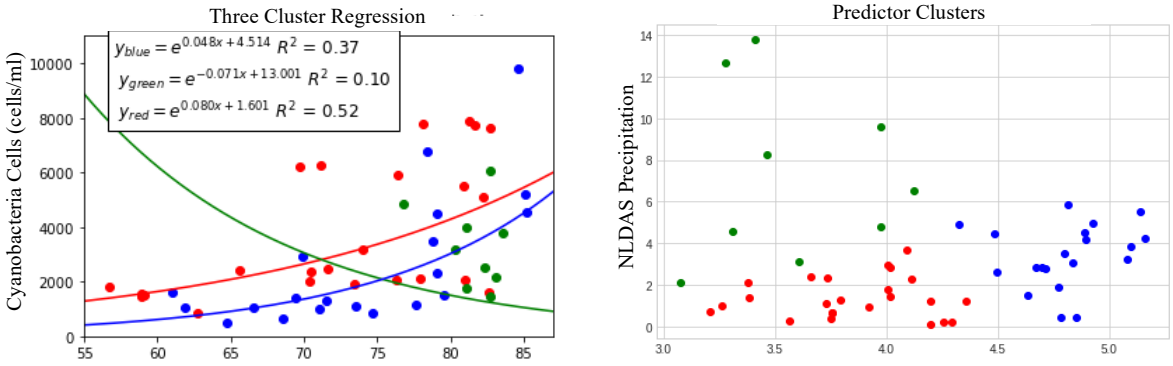

Figure 20. Nonlinear regression with three-cluster regression with outliers included using the NLDAS 5-day windspeed, 14-day shortwave radiation, and 32-day precipitation for clusters. The dependent variable is 3-day cyanobacteria data and the independent variable is MODIS 32-day LST.

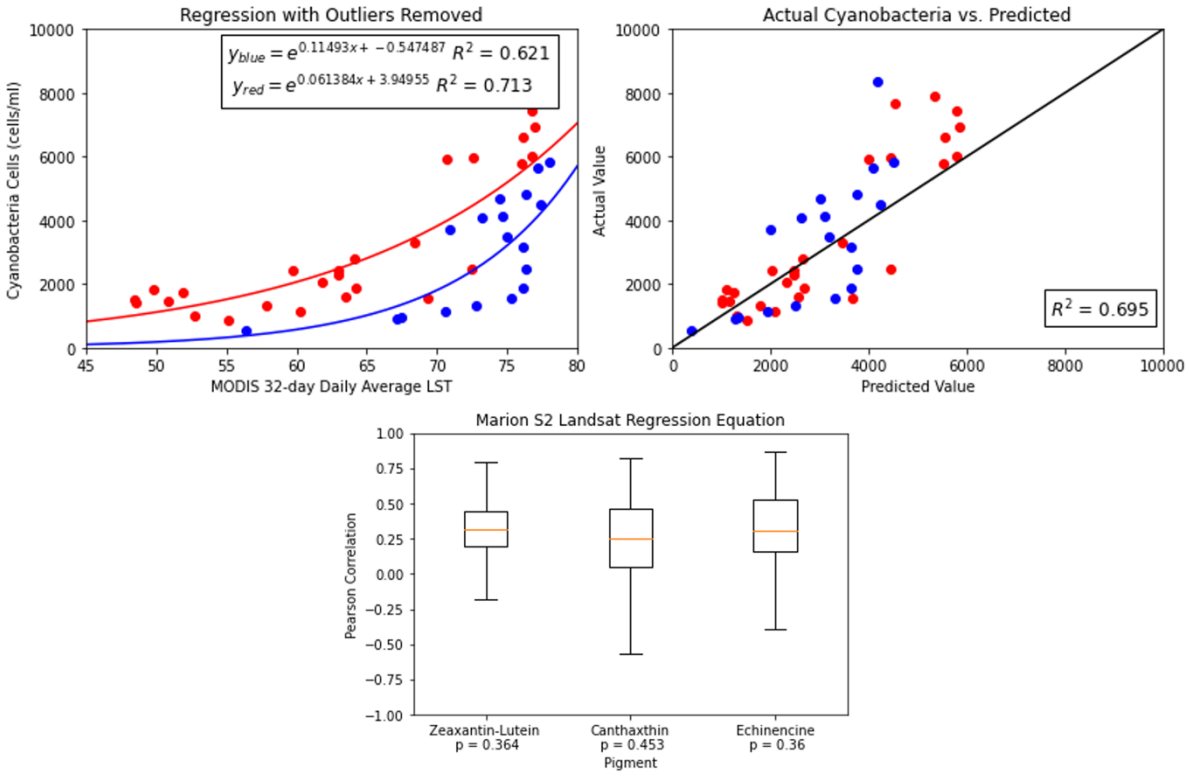

Figure 21. Final nonlinear regression with six outliers removed, using NLDAS 5-day windspeed, 14-day shortwave radiation, and 32-day precipitation in clusters. The dependent variable is cyanobacteria data and the independent variable is MODIS 32-day daily average LST. Box and whisker plots show the median correlations for final MODIS regression equations applied to Marion sediment core pigments. Median p-values are included with each pigment.

outliers was tested with the daily average, daily, and nighttime temperatures, and the daily average temperatures produced the highest R-squared for the actual versus predicted equation, so this temperature was used for the final mode. When the daily average was tested alone with cyanobacteria, it had shown poorer correlation than the daytime values. It may be better for the regression equation though since multiple variables are included to cluster these datasets, which could allow for the more complex relationships between temperature and the other environmental variables to become apparent. It also likely reflects the importance of nighttime temperatures in the process of bloom growth, making the increasing nighttime temperatures an important part of the regression.

Transferability of the model to the sediment core locations is key for application on a statewide basis. The relationship with pigment data and regression predictions was tested with many of the regression equations, typically the correlation was near zero for all locations. The final regression model showed the best transferability to other locations, particularly Marion Reservoir (Figure 21). Marion likely had the strongest relationship because of its similarity to Cheney—both have dams positioned with their outlets facing southeast and the main length of the reservoir is in this direction as well.

LandSat Model: The regression process was then repeated with the LandSat dataset, the best averaging period for regressions was 32 days for windspeed and precipitation, 32-day temperature, and with shortwave radiation being removed (Figure 22). This dataset did not produce

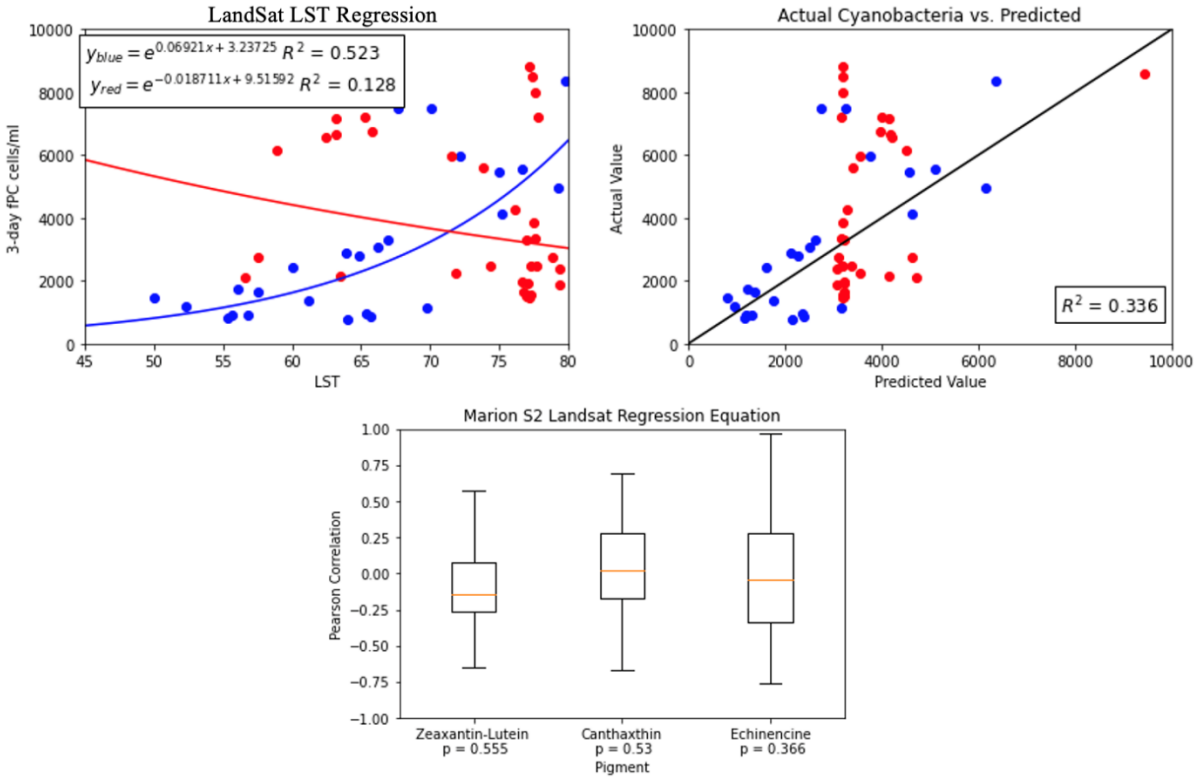

Figure 22. LandSat final regression model with Cheney cyanobacteria. Two separate clusters with inputs of NLDAS 32-day averages of precipitation, windspeed, and shortwave radiation are included, and the regression is modeled using 3-day cyanobacteria data as the independent variable and LandSat 32-day temperature as the dependent variable. The box and whisker plot shows application to the Marion pigment data as median correlations with median p-values listed.

a clear relationship at Cheney nor did it provide strong transferability to Marion, even with the large number of datapoints. Outliers were not removed from this dataset, as no clear points existed. To account for some of the high values that appeared in the middle of the temperature range, a third cluster was added; however, the third cluster only had 9 samples, and did not lead to any improvements in the actual versus predicted correlation. Because of this, the two cluster regressions were used for application at Marion. Regressions with the NDVI were also attempted, but these regressions did not improve over those without it. The NDVI appeared to be redundant with the temperature when using this model as well. The inconsistent temporal frequency of the LandSat dataset, despite the high spatial resolution, may be the culprit of the poor modeling at Cheney leading to the poor results at Marion. While the dataset provided information from multiple satellites with multiple overpasses, some months had as little as one value for the temperature and NDVI. Moreover, the percent coverage of the satellite may not represent the entire reservoir well even when there are multiple values in a short period and produce a temperature and NDVI that are less representative of the reservoir as a whole.

Merged Model: For a larger dataset, LandSat and MODIS temperatures and NDVI were combined. The best averaging period for this dataset was the same as the MODIS regressions: 32-day temperature, 32-day precipitation, 5-day windspeed, and 14-day shortwave radiation (Figure 23). This model produced two clear regressions from distinct clusters but lacked in transferability to the sediment core data at Marion. Adding NDVI also did not lead to improvements. The poor performance here could stem from differences in the two sensors that make them incompatible for combining datasets.

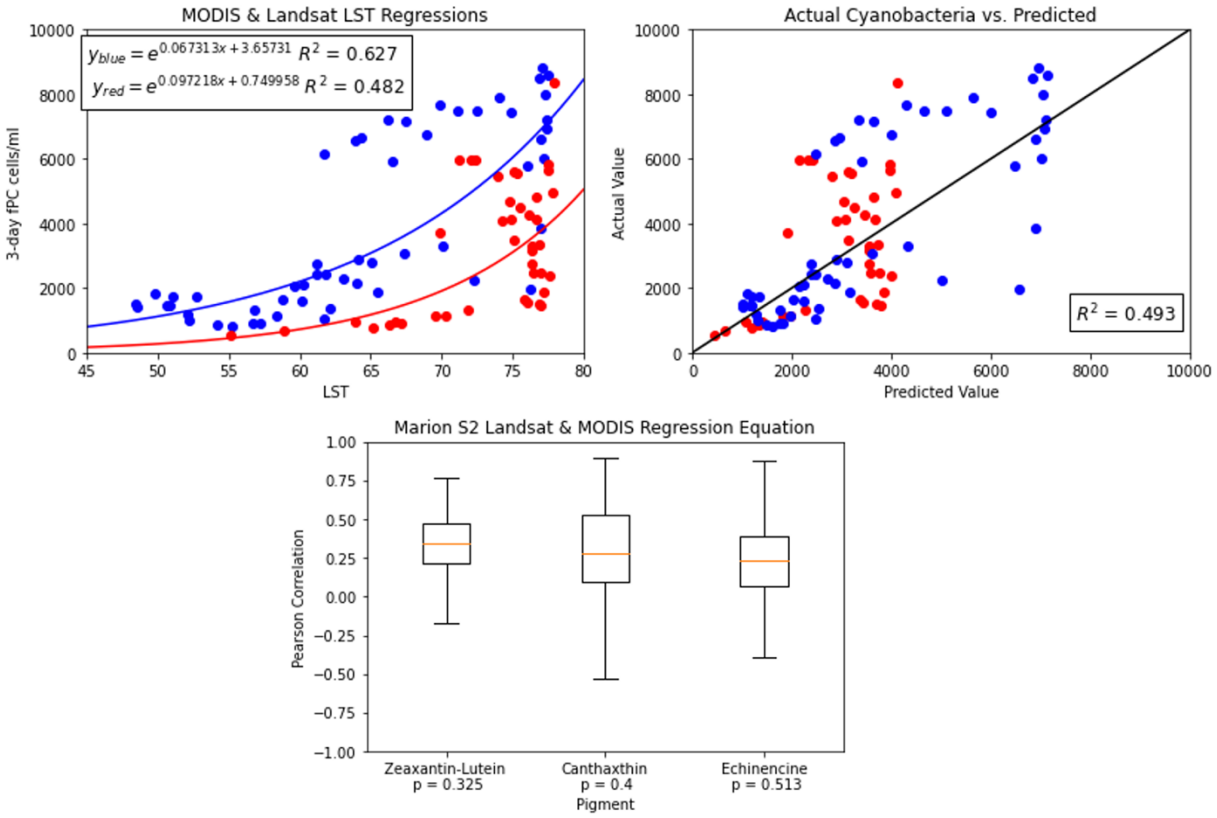

Figure 23. Merged final regression model with Cheney cyanobacteria. Two separate clusters with inputs of NLDAS 5-day windspeed, 14-day shortwave radiation, and 32-day precipitation are included, and the regression is modeled using 3-day cyanobacteria data as the independent variable and Merged 32-day temperature as the dependent variable. The box and whisker plot shows application to the Marion pigment data through median correlation, and median p-values are listed.

3.3.3 Regression Tree Models

Regression trees offered a more complex way of sorting data than clustering and were explored below to search for further improvements in the model's transferability to sediment core data. The regression tree models used the same rolling average periods for the variables used in them as the clustering models, as well as the 3-day average cyanobacteria values. Predictors for each tree and their importance level are shown in Table 3. All final models at Cheney show the application of the decision tree on the combined test and train dataset.

Table 3. Input predictors for each regression tree model and their feature importance.

Model	Predictor	Importance
NLDAS	32-day Precipitation	0.385
	32-day Windspeed	0.255
	32-day Shortwave Radiation	0.360
MODIS	32-day Daily Average Surface Temperature	0.542
	32-day Windspeed	0.091
	32-day Precipitation	0.239
	14-day Shortwave Radiation	0.127
LandSat	32-day NDVI	0.271
	32-day Windspeed	0.071
	14-day Shortwave Radiation	0.658
MODIS + LandSat	32-day Daily Average Surface Temperature	0.254
	32-day NDVI	0.238
	14-day Shortwave Radiation	0.043
	32-day Precipitation	0.464

NLDAS Model: The NLDAS model, having the largest dataset, also had the largest regression tree with 68 leaves for potential prediction (Figure 24). The model created by this tree had a correlation of 0.95 but when transferred to the sediment core data at Marion, the median correlation was low (Figure 25). Canthaxthin had the strongest relationship when it was applied, and zeaxanthin-lutein was similarly related. Because of the many paths for relating the variables, this model may be

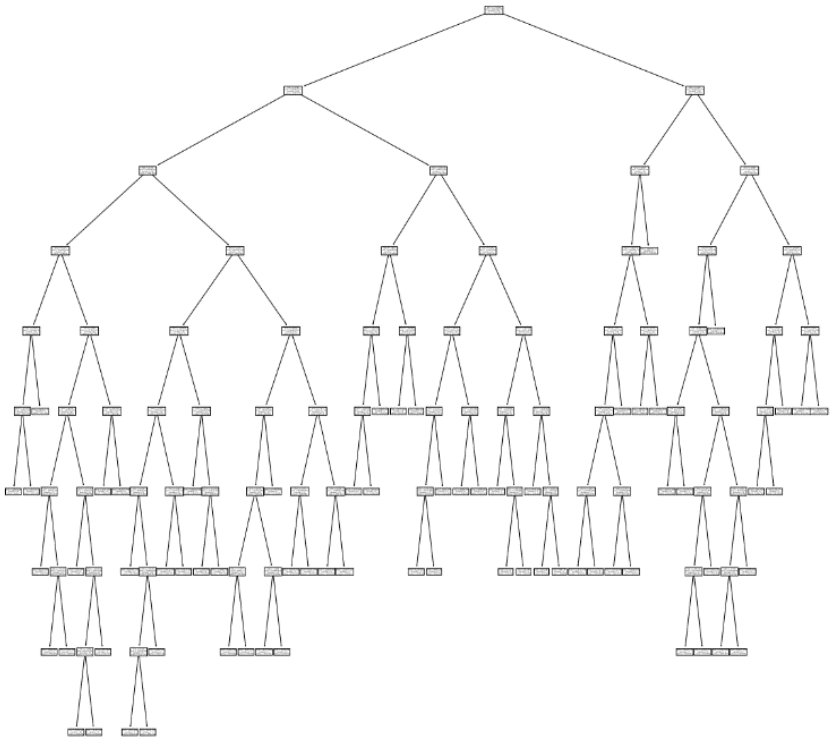

Figure 24. NLDAS regression tree with initial node using the NLDAS 32-day

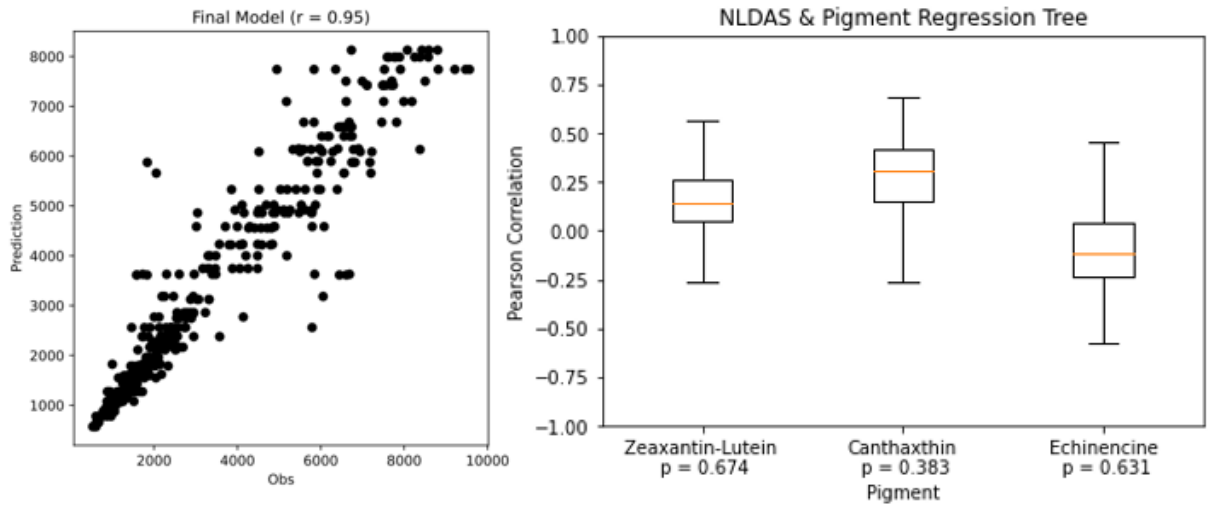

Figure 25. NLDAS regression tree predictions at Cheney (left) and a box and whisker plot of the model applied at Marion, showing median correlations, and listing median p-values.

overfitting to the dataset at Cheney Lake and taking away from the ability to use it at other locations.

MODIS Model: The MODIS model had the smallest decision tree from the smallest remote sensing dataset, with only 6 leaves (Figure 26). The regression model had a correlation of 0.79 and showed decent transferability to canthaxthin at Marion (Figure 27). This suggests that the model may not be as over calibrated for Cheney, but it indicates that there may be less transferability between the Cheney and Marion.

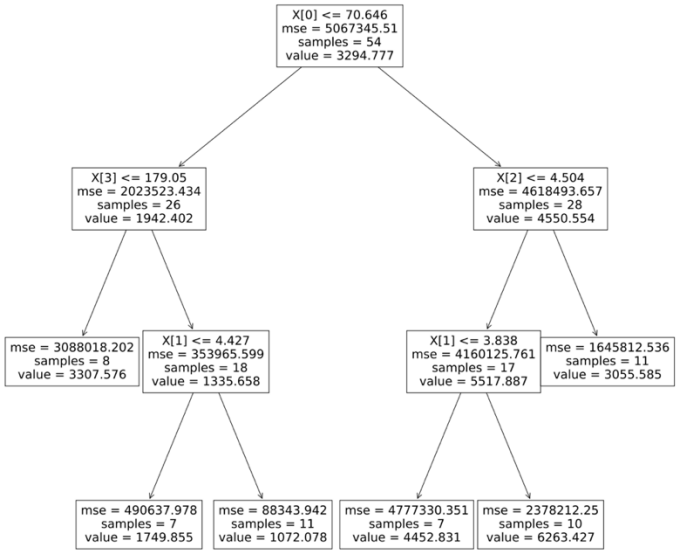

Figure 26. MODIS regression tree with the initial node using the MODIS 32-day average temperature.

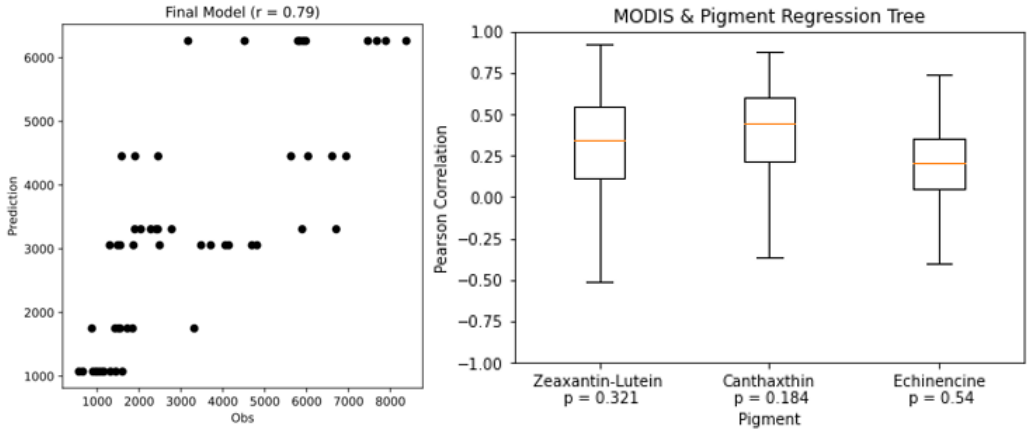

Figure 28. MODIS regression tree predictions at Cheney (left) and a box and whisker plot of the model applied at Marion, showing median correlations, and listing median p-values.

LandSat Model: The LandSat model had 10 leaves in the regression tree (Figure 28). The regression model provided very similar outputs for Marion as the clustering model, showing nearly no improvement for predictions when applied with the sediment core data (Figure 29). The issue

Figure 27. LandSat regression tree model, where the initial node is the 32-day LandSat NDVI.

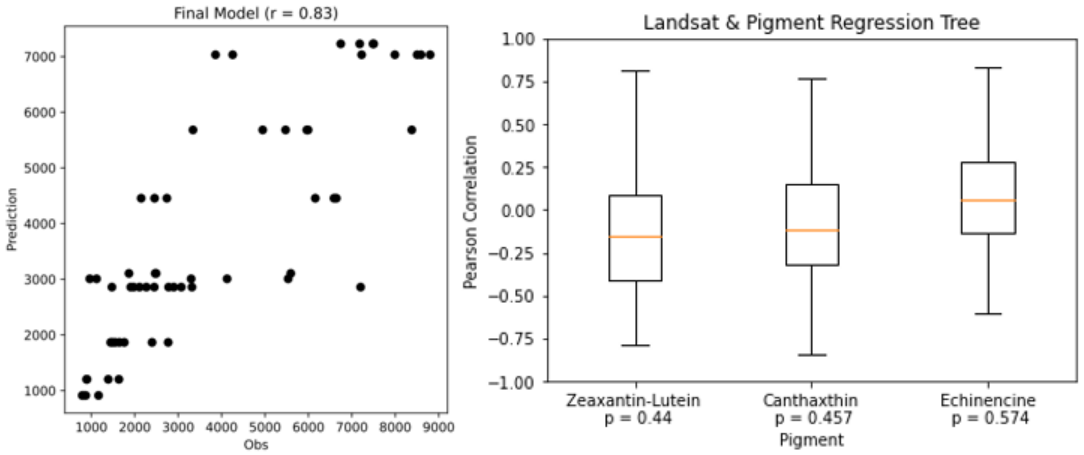

Figure 29. LandSat regression tree predictions at Cheney (left) and a box and whisker plot of the model applied at Marion, showing median correlations and listing median p-values.

with applying LandSat datasets here may be not only because of over calibration but also again the inconsistent temporal steps in the dataset may not be capturing the relationship well with the surface temperature.

Merged Model: The decision tree for the merged model had 13 leaves for possible ending points of the model (Figure 30). The model performed well, with the highest correlations of all regression models; moreover, while it did not do well with the sediment core data, it improved over LandSat alone, but not MODIS (Figure 31). Again, the concern of over calibrating comes up for this model. The lack of consistency in timesteps is somewhat resolved by combining the two satellite datasets, but it does

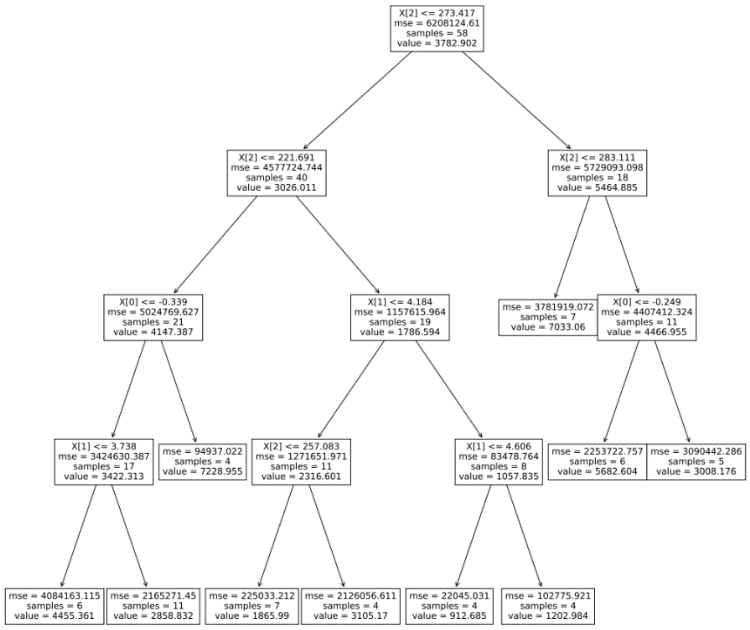

Figure 30. Merged regression tree model, where the initial node is the 32-day combined LST.

not cause improvements over MODIS alone, meaning that for some reason LandSat is degrading the MODIS based prediction.

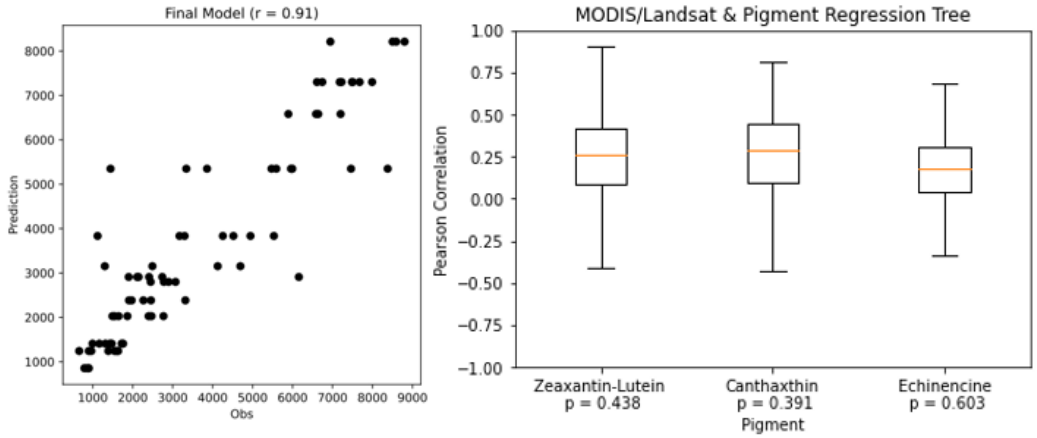

Figure 31. Merged regression tree predictions at Cheney (left) and a box and whisker plot of the model applied at Marion, showing median correlations and listing median p-values.

Chapter 4: Discussion and Conclusions

4.1 Summary

Table 4. Model correlations for nonlinear regressions and regression trees

Dataset	Cheney		Marion	
	Nonlinear Regression	Regression Tree	Nonlinear Regression (Zeaxantin-Lutein/ Canthaxthin/ Echinecine)	Regression Tree (Zeaxantin-Lutein/ Canthaxthin/ Echinecine)
MODIS	0.704	0.790	0.321 0.236 0.312	0.334 0.442 0.210
LandSat	0.366	0.830	-0.140 0.026 -0.043	-0.153 -0.115 0.059
MODIS + LandSat	0.493	0.910	0.345 0.277 0.231	0.264 0.286 0.177
NLDAS	0.445	0.950	0.171 0.312 0.132	0.145 0.307 -0.113

Looking at the first question of what trends and relationships exist between environmental variables at Kansas reservoirs, temperatures are increasing in the summer months and decreasing in May. Moreover, EVI and NDVI were found to be related to NLDAS parameter's variability, while the FAI had a more unique source of variability from the NLDAS data. The second question, which considers the key drivers of HAB events in Kansas reservoirs, resulted in finding that temperature, windspeed, shortwave radiation, and precipitation had the strongest relationship with cyanobacteria data at Cheney. These remote sensing variables are related to blooms and could be drivers or be second drivers of HABs if they are reflecting the seasonality of known drivers of HABs. Nutrients entering reservoirs is seasonal, since crops are grown seasonally, so the remote sensing data may be capturing this. Moreover, precipitation specifically could be reflecting the relationship of increased blooms with decreased sediment concentrations. The sediment load after

a precipitation event would be higher from increased erosion, and then settling out over time, which could be the phenomena captured by the relationship with precipitation. Ultimately, the remote sensing and reanalysis variables are potentially capturing multiple facets of what is happening in the reservoir.

The last question was about this work's major goal to find satellite predictors of CyanoHABs in Kansas reservoirs and use that information to model blooms using satellite data. Of the variables explored, surface temperature, windspeed, shortwave radiation, and precipitation showed the strongest relationship with the cyanobacteria at Cheney Reservoir and more generally with the seasonal trends. Table 3 shows the results of all final models for different datasets found in the study. From these, a nonlinear model showed to have the most power when transferred from the model creation location to another reservoir. The regression tree, despite the ability to characterize more specific groups of data, performed better at the initial location, but did not produce that same transferability. The downfall of these model may have been overfitting the model to the first location, making it too specific to be applied elsewhere.

Between MODIS, LandSat, and a combination of the two, MODIS performed the best at Cheney with the nonlinear regression and the combined dataset performed the best here with the regression tree. Moreover, when the MODIS nonlinear regression and regression tree models were applied and compared to the sediment core data for Marion reservoir, it proved to have the most transferability of the models from satellite data, though it was not perfect. MODIS may have performed better due to its own strengths but also issues within the other datasets. First, MODIS has the most consistent and frequent timestep. However, combining the two datasets should have remedied issue of frequency in the LandSat dataset, but it did not. This leads to the next potential issue with LandSat data: while it has a finer spatial resolution than the MODIS data, it could be

degrading the prediction by capturing more spatial variability in the lake, being less representative. This was explored by using a smaller, more centralized image of LandSat, but it did not improve the prediction. This could also be due to the LandSat data having much more noise than MODIS. When processing the LandSat data, images with at least 10% coverage after removing cloud effected pixels were used in the analysis. Cloud pixels were identified by using pixels classified with low confidence cloud. It could be that this coverage threshold is too low and high cloud cover days in conjunction with inconsistent temporal resolution may be the culprit of the degradation of LandSat data in this study.

Moving forward, using MODIS with a nonlinear, clustered approach is a strong starting point for other uses. LandSat should not be completely discounted though, as combining the models does show some potential, at least with regression trees, to provide predictability. It may prove more useful in modeling individual reservoirs but not as helpful when looking for a model that can easily be adjusted to other locations.

4.2 Uncertainty and Limitations

A limitation with this study is the temporal resolution of the satellite data, which is important for making management decisions (Papenfus et al. 2020). Since MODIS only provides data every 8 days, there may be additional indictors in this data that are not well-represented in the dataset due to this. It also only captures information about the surface of the water and blooms extend into the water column. Moreover, this creates a smaller overall sample size for analysis. LandSat typically does have a higher temporal resolution, but the inconsistency in overpasses often leads to lower temporal resolution in some months. Combining the two somewhat remedies the issue of temporal resolution, but still does not lead to improvements over the period analyzed. Another limitation is that there are many hydrological, biological, and physiochemical drivers of

CyanoHABs which cannot all be accounted for in this project. The goal of this prediction framework is to identify if satellite data contains potential indicators of the blooms. Future work assimilating satellite data into the General Lake Model should include more of the known drivers of CyanoHABs while the satellite data adds onto it to create a more accurate model.

Another limitation is the lack of in situ data at many reservoirs in Kansas to use for validation. The Cheney Reservoir datasets is helpful but is limited to only a few years. The sediment core data is useful in this analysis, but it is simply too coarse to make short-term predictions. The regression is only applicable to Cheney Reservoir due to this lack of data for other reservoirs. It is also limited to the temperature range the regression was completed with and limited to the summer. Few values were available in the winter and algal blooms are less of a concern in the colder months in Kansas. Additionally, the variation of the depth of the sensor causes unreliability of the values collected in terms of representing a constant depth near the surface, since that would be most represented by the remote sensing.

For Cheney, an additional dataset exists from late 2014 to present, after the USGS replaced the first sensor with a newer one. This sensor outputs measurements in relative fluorescence units (RFU) rather than cells per milliliter like the older instrument. The older sensor has a converts RFU to cells per milliliter by multiplying by a factor of about 2,800 (YSI 2006). The two sensors use different optical wavelengths for measurements and are highly variable, so there are no documents from USGS comparing the two. Because of the uncertainty between these sensors they are primarily only used qualitatively. Analysis of the overlapping data from the two confirmed that the two datasets could do not consistently represent similar values. Attempts to use this data were made, but essentially no relationship was found between any of the parameters used and the data from this newer sensor.

4.3 Impacts and Future Work

While the model provided a transferrable model from Cheney to Marion—two similar reservoirs—the issue of application at other reservoirs still exists. Being able to apply a model on a more widespread basis throughout the state would provide a more powerful predictive model. This research provides a foundation for future work in providing assimilation inputs for lake models and helping better understand how blooms can be predicted. The General Lake Model, a one-dimensional model for simulating lake hydrodynamics and ecological dynamics via the coupled AquaticEcoDyamics, driven by meteorological and lake water inflow/outflow data (Hipsey et al. 2019) could be a potential solution. Assimilating the remote sensing data into this model will further develop the relationship between the satellite drivers found in this paper to create a stronger model with the data that could provide a more robust prediction framework. Further work into creating stronger models should include additional machine learning techniques on larger datasets to better represent the dynamic relationship between cyanobacteria and remote sensing data for assimilation. Rectifying the lack of validation data will be the major hurdle for completing this work and making a more useful prediction framework.

References

Anderson, Donald M., Porter Hoagland, Yoshi Kaoru, and Alan W. White. 2000. *Estimated Annual Economic Impacts from Harmful Algal Blooms (HABs) in the United States*. Woods Hole, MA: Woods Hole Oceanographic Institution. https://doi.org/10.1575/1912/96.

Araya, A., I. Kisekka, X. Lin, P. V. Vara Prasad, P. H. Gowda, C. Rice, and A. Andales. 2017. "Evaluating the Impact of Future Climate Change on Irrigated Maize Production in Kansas." *Climate Risk Management* 17 (February): 139–54. https://doi.org/10.1016/j.crm.2017.08.001.

Brikowski, T. H. 2008. "Doomed Reservoirs in Kansas, USA? Climate Change and Groundwater Mining on the Great Plains Lead to Unsustainable Surface Water Storage." *Journal of Hydrology* 354 (1–4): 90–101. https://doi.org/10.1016/j.jhydrol.2008.02.020.

Camps-Valls, Gustau, Manuel Campos-Taberner, Álvaro Moreno-Martínez, Sophia Walther, Grégory Duveiller, Alessandro Cescatti, Miguel D. Mahecha, et al. 2021. "A Unified Vegetation Index for Quantifying the Terrestrial Biosphere." *Science Advances* 7 (9): eabc7447. https://doi.org/10.1126/sciadv.abc7447.

Carmichael, Wayne W. 2001. "Health Effects of Toxin-Producing Cyanobacteria: 'The CyanoHABs.'" *Human and Ecological Risk Assessment: An International Journal* 7 (5): 1393–1407. https://doi.org/10.1080/20018091095087.

Christensen, Victoria G, Jennifer L. Graham, Chad R. Milligan, Larry M. Pope, and Andrew C. Ziegler. 2006. "Water Quality and Relation to Taste-and-Odor Compounds in the North Fork Ninnescah River and Cheney Reservoir South-Central Kansas, 1997-2003." *U.S. Geological Survey*, Scientific Investigations Report, , no. 5095.

Coffer, M. M., B. A. Schaeffer, J. A. Darling, E. A. Urquhart, and W. B. Salls. 2020. "Quantifying National and Regional Cyanobacterial Occurrence in US Lakes Using Satellite Remote Sensing." *Ecol Indic* 111 (April): 105976. https://doi.org/10.1016/j.ecolind.2019.105976.

Didan, K. 2015a. "MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006." *NASA EOSDIS Land Processes DAAC.* https://doi.org/10.5067/MODIS/MOD13Q1.006.

———. 2015b. "MYD13A1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 500m SIN Grid V006." *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MYD13A1.006.

Didan, Kamel, Armando Barreto Munoz, and Alfredo Huete. 2015. "MODIS Vegetation Index User's Guide (MOD13 Series)," June, 35.

Downing, John A, Susan B Watson, and Edward McCauley. 2001. "Predicting Cyanobacteria Dominance in Lakes." *Canadian Journal of Fisheries and Aquatic Sciences* 58 (10): 1905–8. https://doi.org/10.1139/f01-143.

Dwyer, John L., David P. Roy, Brian Sauer, Calli B. Jenkerson, Hankui K. Zhang, and Leo Lymburner. 2018. "Analysis Ready Data: Enabling Analysis of the Landsat Archive." *Remote Sensing* 10 (9): 1363. https://doi.org/10.3390/rs10091363.

Dzialowski, Andrew R., Donald G. Huggins, Jerry deNoyelles, Mark Jakubauskas, Niang Choo Lim, and Jason Beury. 2008. "Predicting Taste and Odor Events in Kansas Reservoirs." June 2008. https://documentcloud.adobe.com/spodintegration/index.html?r=1&locale=en-us.

Graham, Jennifer L., Guy M. Foster, and Ariele R. Kramer. 2017. “Twenty Years of Water-Quality Studies in the Cheney Reservior Watershed, Kansas, 1996-2016.” Fact Sheet. Fact Sheet.

Graham, Jennifer L., John R. Jones, Susan B. Jones, John A. Downing, and Thomas E. Clevenger. 2004. “Environmental Factors Influencing Microcystin Distribution and Concentration in the Midwestern United States.” 2004. https://doi.org/10.1016/j.watres.2004.08.004.

Graham, Jennifer L., Keith A. Loftin, Michael T. Meyer, and Andrew C. Ziegler. 2010. “Cyanotoxin Mixtures and Taste-and-Odor Compounds in Cyanobacterial Blooms from the Midwestern United States.” *Environmental Science & Technology* 44 (19): 7361–68. https://doi.org/10.1021/es1008938.

Graham, Jennifer L., Andrew C. Ziegler, B. L. Loving, and Keith A. Loftin. 2012. “Fate and Transport of Cyanobacteria and Associated Toxins and Taste-and-Odor Compounds from Upstream Reservoir Releases in the Kansas River, Kansas, September and October 2011.” Scientific Investigations Report. Scientific Investigations Report.

Harris, Ted, Debra S. Baker, Jennifer Moody, Jude Kastens, Belinda Sturm, Peter Leavitt, and Michael Ketterer. 2021. “Phytoplankton and Water Quality in Kanopolis and Webster Reservoirs: Results of Paleolimnological Sediment Core and Historical Data Analyses.” *Kansas Biological Survey* 202 (May).

Harris, Ted, Jin-Ho Yun, Debra S. Baker, Jude Kastens, Belinda Sturm, Peter Leavitt, Michael Ketterer, and Ann St. Amand. 2020a. “Phytoplankton and Water Quality in Marion and Keith Sebelius Reservoirs: Results of Paleolimnological Sediment Core and Historical Data Analyses.” *Kansas Biological Survey* 198 (February).

———. 2020b. “Phytoplankton and Water Quality in Milford Reservoir: Results of Paleolimnological Sediment Core and Historical Data Analyses.” *Kansas Biological Survey* 197 (February).

Hipsey, Matthew R., Louise C. Bruce, Casper Boon, Brendan Busch, Cayelan C. Carey, David P. Hamilton, Paul C. Hanson, et al. 2019. “A General Lake Model (GLM 3.0) for Linking with High-Frequency Sensor Data from the Global Lake Ecological Observatory Network (GLEON).” *Geoscientific Model Development* 12 (1): 473–523. https://doi.org/10.5194/gmd-12-473-2019.

Hu, Chuanmin. 2009. “A Novel Ocean Color Index to Detect Floating Algae in the Global Oceans.” *Remote Sensing of Environment* 113 (10): 2118–29. https://doi.org/10.1016/j.rse.2009.05.012.

Huisman, Jef, H.C.P. Matthijs, and Petra M. Visser. 2005. “Harmful Cyanobacteria.” *Aquatic Ecology Series 3, Dordrecht, the Netherlands: Springer*.

Huntington, Thomas G. 2006. “Evidence for Intensification of the Global Water Cycle: Review and Synthesis.” *Journal of Hydrology* 319 (1–4): 83–95. https://doi.org/10.1016/j.jhydrol.2005.07.003.

Huppert, Amit, Bernd Blasius, Ronen Olinky, and Lewi Stone. 2005. “A Model for Seasonal Phytoplankton Blooms.” *Journal of Theoretical Biology* 236 (May): 276–90. https://doi.org/doi:10.1015/j.jtbi.2005.03.012.

Jiang, Z, A Huete, K Didan, and T Miura. 2008. “Development of a Two-Band Enhanced Vegetation Index without a Blue Band.” *Remote Sensing of Environment* 112 (10): 3833–45. https://doi.org/10.1016/j.rse.2008.06.006.

KDHE. n.d. “Blue-Green Algae (BGA) Blooms.” Accessed September 18, 2021a. http://www.kdheks.gov/algae-illness/.

———. n.d. “What Do the Different Terms Mean?” Kansas Department of Health and Environment. Accessed June 4, 2022b. https://www.kdhe.ks.gov/FAQ.aspx?QID=436.

Leavitt, Peter R., and Dominic A. Hodgson. 2002. “Sedimentary Pigments.” In *Tracking Environmental Change Using Lake Sediments*, edited by John P. Smol, H. John B. Birks, William M. Last, Raymond S. Bradley, and Keith Alverson, 3:295–325. Developments in Paleoenvironmental Research. Dordrecht: Springer Netherlands. https://doi.org/10.1007/0-306-47668-1_15.

Lin, X., J. Harrington, I. Ciampitti, P. Gowda, D. Brown, and I. Kisekka. 2017. “Kansas Trends and Changes in Temperature, Precipitation, Drought, and Frost-Free Days from the 1890s to 2015.” *Journal of Contemporary Water Research & Education* 162 (1): 18–30. https://doi.org/10.1111/j.1936-704X.2017.03257.x.

Mahmud, M. Hussain and I. n.d. “PyMannKendall: A Python Package for Non Parametric Mann Kendall Family of Trend Tests. | EndNote Click.” Accessed April 9, 2022. https://click.endnote.com/viewer?doi=10.21105%2Fjoss.01556&token=WzM2MDQxMTQsIjEwLjIxMTA1L2pvc3MuMDE1NTYiXQ.B-b5kCK_q4-eaCYo73hB2XvE87Q.

Mann, Henry B. 1945. “Nonparametric Tests Against Trend.” *Econometrica* 13 (3): 245–59. https://doi.org/10.2307/1907187.

McCabe, Gregory J, and David M Wolock. 2011. “Independent Effects of Temperature and Precipitation on Modeled Runoff in the Conterminous United States.” *Water Resources Research* 47 (11): W11522. https://doi.org/10.1029/2011wr010630.

McClain, Charles R. 2009. “A Decade of Satellite Ocean Color Observations.” *Annual Review of Marine Science* 1 (1): 19–42. https://doi.org/10.1146/annurev.marine.010908.163650.

Mouw, Colleen B., Steven Greb, Dirk Aurin, Paul M. DiGiacomo, Zhongping Lee, Michael Twardowski, Caren Binding, et al. 2015. “Aquatic Color Radiometry Remote Sensing of Coastal and Inland Waters: Challenges and Recommendations for Future Satellite Missions.” *Remote Sensing of Environment* 160: 15–30. https://doi.org/10.1016/j.rse.2015.02.001.

Murphy, E. B., K. A. Steidinger, B. S. Roberts, Jerome Williams, and E.B. Murphy. 1975. “An Explanation for the Florida East Coast Gymnodinium Breve Red Tide of November 1972.” *Limnology and Oceanography*, no. 1975 vol. 20 n° 3: 481–86. https://doi.org/doi:https://doi.org/10.4319/lo.1975.20.3.0481.

NOAA National Centers for Environmental Information. 2020. “State of the Climate: Monthly Global Climate Report for Annual 2020.” https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202013.

Paerl, Hans W., and Jef Huisman. 2008. “Blooms Like It Hot.” *Science* 320 (5872): 57–58. https://doi.org/10.1126/science.1155398.

———. 2009. “Climate Change: A Catalyst for Global Expansion of Harmful Cyanobacterial Blooms.” *Environmental Microbiology Reports* 1 (1): 27–37. https://doi.org/10.1111/j.1758-2229.2008.00004.x.

Paerl, Hans W., and Timothy G. Otten. 2012. “Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls,” December. https://click.endnote.com/viewer?doi=10.1007%2Fs00248-012-0159-y&token=WzM2MDQxMTQsIjEwLjEwMDcvczAwMjQ4LTAxMi0wMTU5LXkiXQ.HimMqKvY1XKih5Sf-kyEng2MDjo.

Papenfus, Michael, Blake Schaeffer, Amina I. Pollard, and Keith Loftin. 2020. “Exploring the Potential Value of Satellite Remote Sensing to Monitor Chlorophyll-a for US Lakes and Reservoirs.” *Environmental Monitoring and Assessment* 192 (12): 808. https://doi.org/10.1007/s10661-020-08631-5.

“Remote Sensing: An Overview | Earthdata.” n.d. Accessed April 29, 2022. https://earthdata.nasa.gov/learn/backgrounders/remote-sensing/.

Seabold, Skipper, and Josef Perktold. 2010. “Statsmodels: Econometric and Statistical Modeling with Python.” In . Proceedings of the 9th Python in Science Conference. https://www.statsmodels.org/stable/index.html.

Shang, Xingxing, Xiaohui Jiang, Ruining Jia, and Chen Wei. 2019. “Land Use and Climate Change Effects on Surface Runoff Variations in the Upper Heihe River Basin.” *Water* 11 (2): 344. https://doi.org/10.3390/w11020344.

Stroming, Signe, Molly Robertson, Bethany Mabee, Yusuke Kuwayama, and Blake Schaeffer. 2020. “Quantifying the Human Health Benefits of Using Satellite Information to Detect Cyanobacterial Harmful Algal Blooms and Manage Recreational Advisories in U.S. Lakes.” *GeoHealth* 4 (9): e2020GH000254. https://doi.org/10.1029/2020GH000254.

Stumpf, R. P., and D. Dupuy. 2016. “Experimental Lake Erie Harmful Algal Bloom Bulletin.” *NOAA*, 2016, sec. Bulletin 28.

Tonk, Linda, Kim Bosch, Petra M. Visser, and Jef Huisman. 2007. “Salt Tolerance of the Harmful Cyanobacterium Microcystis Aeruginosa.” *Aquatic Microbial Ecology* 46: 117–23.

US EPA, ORD. 2014. “Cyanobacteria Assessment Network (CyAN).” Overviews and Factsheets. June 18, 2014. https://www.epa.gov/water-research/cyanobacteria-assessment-network-cyan.

———. 2017. “Fresh Surface Water.” Reports and Assessments. November 2, 2017. https://www.epa.gov/report-environment/fresh-surface-water.

U.S. Geological Survey. 2022. “National Water Information System Data Available on the World Wide Web (Water Data for the Nation).” U.S. Geological Survey. 2022. https://waterdata.usgs.gov/ks/nwis/uv/?site_no=07144790.

Vermonte, E. 2015a. “MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V006.” *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MOD09A1.006.

———. 2015b. “MYD09A1 MODIS/Aqua Surface Reflectance 8-Day L3 Global 500m SIN Grid V006.” *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MYD09A1.006.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Wan, Z., S. Hook, and G. Hulley. 2015a. “MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006 [Data Set].” *NASA EOSDIS Land Processes DAAC.* https://doi.org/10.5067/MODIS/MOD11A2.006.

———. 2015b. “MYD11A2 MODIS/Aqua Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. 2015.” *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MYD11A2.006.

Wang, Dingbao, and Mohamad Hejazi. 2011. “Quantifying the Relative Contribution of the Climate and Direct Human Impacts on Mean Annual Streamflow in the Contiguous United States.” *Water Resources Research* 47 (9): W00J12. https://doi.org/10.1029/2010wr010283.

Wulder, Michael A., and Nicholas C. Coops. 2014. “Satellites: Make Earth Observations Open Access.” *Nature* 513 (7516): 30–31. https://doi.org/10.1038/513030a.

Xia, Y., K. Mitchell, M. Ek, J. Sheffield, B. Cosgrove, E. Wood, L. Luo, et al. 2012. “Continental-Scale Water and Energy Flux Analysis and Validation for the North American Land Data Assimilation System Project Phase 2 (NLDAS-2): 1. Intercomparison and Application of Model Products.” *Journal of Geophysical Research* 117. https://doi.org/10.1029/2011JD016048.

YSI. 2006. “YSI 6131 and 6132 Blue-Green Algae Sensors.” YSI.