UTILIZING GEOSPATIAL CLOUD COMPUTING AND DATA ANALYTICS FOR 
CYANOBACTERIA HARMFUL ALGAL BLOOM RISK MAPPING IN GEORGIA 
PIEDMONT WATERBODIES

by
SAMUEL JOSEPH WEBER
(Under the Direction of Deepak Mishra)

ABSTRACT

The frequency and severity of cyanobacteria harmful blooms (CyanoHABs) have been increasing, with eutrophication and shifting climate paradigms being identified as primary culprits. CyanoHABs produce a spectrum of toxins and can trigger neurological disorder, organ failure, and even death. To promote proactive CyanoHAB management, geospatial risk modeling can act as a predictive mechanism to supplement current mitigation efforts. In this study, exploratory data analysis techniques were used to identify the strongest CyanoHAB predictors based on Sentinel 2A-derived cyanobacteria cell densities for 771 waterbodies in the Georgia Piedmont, and watershed landscape characteristics utilizing Google Earth Engine. Watershed maximum winter temperature, percent agriculture, percent forest, percent impervious, and waterbody area were the strongest predictors of cyanobacteria cell density with a 0.33 R-squared. A Jenks Natural Break scheme assigned waterbodies to CyanoHAB risk groups, and of the 771 waterbodies, 24.38% were low, 37.35% were medium, and 38.26% high risk respectively.

INDEX WORDS: Cyanobacteria, Watershed Modeling, Sentinel-2A, Google Earth Engine, Land Use, Land Cover, Risk Mapping, NDCI
UTILIZING GEOSPATIAL CLOUD COMPUTING AND DATA ANALYTICS FOR CYANOBACTERIA HARMFUL ALGAL BLOOM RISK MAPPING IN GEORGIA PIEDMONT WATERBODIES

by

SAMUEL JOSEPH WEBER

B.S., Virginia Tech, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2017
UTILIZING GEOSPATIAL CLOUD COMPUTING AND DATA ANALYTICS FOR CYANOBACTERIA HARMFUL ALGAL BLOOM RISK MAPPING IN GEORGIA PIEDMONT WATERBODIES

by

SAMUEL JOSEPH WEBER

Major Professor: Deepak Mishra

Committee: Susan Wilde
Elizabeth Kramer
Marguerite Madden

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2017
ACKNOWLEDGEMENTS

I would like to thank my main advisor, Dr. Deepak Mishra (Geography), and the rest of my committee members; Dr. Susan Wilde (Warnell School of Forestry and Natural Resources), Dr. Elizabeth Kramer (Agricultural and Applied Economics), and Dr. Marguerite Madden (Geography). This project would not have been possible without their diligent advising, constant encouragement, and flexibility. I would also like to thank the United States Department of Agriculture (USDA) Food and Agricultural Sciences National Needs Graduate and Postgraduate Fellowship (NNF) Grants Program for supporting me throughout this project and offering the opportunity to gain valuable training and hands-on work experience. I would finally like to thank my many colleagues, classmates, and friends who have helped me at various points during this project, especially Benjamin Page and Abhishek Kumar.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.01 Analysis of Landscape Trigger Variables</td>
<td>6</td>
</tr>
<tr>
<td>1.02 Google Earth Engine as a Geospatial Processing Tool</td>
<td>10</td>
</tr>
<tr>
<td>1.03 Interdisciplinarity of CyanoHABs in Sustainable Food Systems</td>
<td>11</td>
</tr>
<tr>
<td>2 MATERIALS AND METHODS</td>
<td>16</td>
</tr>
<tr>
<td>2.01 Study Area</td>
<td>16</td>
</tr>
<tr>
<td>2.02 Watershed Delineation Automation</td>
<td>17</td>
</tr>
<tr>
<td>2.03 NAIP LULC Classification</td>
<td>22</td>
</tr>
<tr>
<td>2.04 Google Earth Engine Processing</td>
<td>25</td>
</tr>
<tr>
<td>2.05 Description of Google Earth Engine Datasets</td>
<td>25</td>
</tr>
<tr>
<td>2.06 Cyanobacteria Cell Density Regression</td>
<td>33</td>
</tr>
<tr>
<td>2.07 Cyanobacteria Cell Density Risk Assignment</td>
<td>37</td>
</tr>
<tr>
<td>3 RESULTS AND DISCUSSION</td>
<td>38</td>
</tr>
<tr>
<td>3.01 NAIP and NLCD Accuracy Assessment</td>
<td>38</td>
</tr>
<tr>
<td>3.02 Cyanobacteria Cell Density Risk Map</td>
<td>43</td>
</tr>
<tr>
<td>3.03 Cyanobacteria-Positive Waterbody Validation</td>
<td>58</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: List of acronyms used in this document. .................................................................15
Table 2: Grouping of level II classification classes into level I classification classes. ...........24
Table 3: List of the explanatory variables used in the CCD analysis. All variables are measured at the watershed level, except for waterbody area. .........................................................26
Table 4: Example table of the seven highest CCD values (total study n = 771). .................34
Table 5: Classification accuracy breakdown for the NLCD (A) and NAIP (B) LULC maps. All NAIP LULC classes had a higher consumer’s accuracy than their respective NLCD counterparts, and overall the NAIP was 19.27% more accurate.........................................................41
Table 6: Final MLR outputs for (A) the NLCD LULC based dataset and (B) the NAIP LULC based dataset. The final sample size values (n) represent the original 771 waterbodies with the outliers, defined by over ten times the mean Cook’s distance, omitted. .................48
LIST OF FIGURES

Figure 1: Colored coded map of the 771 waterbodies used in this study by size category. The Georgia Piedmont is outlined by a thick black line. ..........................................................18

Figure 2: Single color map of the 771 waterbodies used in this study. The Georgia Piedmont is outlined by a thick black line. The background image is a Sentinel 2A satellite image from July 26th, 2016. This is the image that was used to calculate CCD for the waterbodies, which is why waterbodies west of the image in the Georgia Piedmont were not included in this study. ..................................................................................................19

Figure 3: Workflow for watershed delineation automation. This process was carried out using ArcGIS ModelBuilder.................................................................21

Figure 4: Climatology histogram for the Georgia Piedmont based on GRIDMET daily high temperature data from 1980 – 2015. The daily images for each month of each year during the 35-year period were averaged to make one average maximum temperature raster per month, and like-months were averaged to create their respective histogram. .................28

Figure 5: Cyanobacteria-positive locations in the Georgia Piedmont confirmed by in-situ sampling by Dr. Susan Wilde from the University of Georgia........................................36

Figure 6: Visual comparison of level I 30-meter NLCD (A, C) and one-meter NAIP (B, D) LULC classifications. A and B are from the same extent, as are C and D. Water is blue, developed is red, forest is green, agriculture is yellow, and wetland is cyan. The images are semi-transparent to allow satellite imagery beneath to be visible... .................................39
Figure 7: Randomly generated points that were used for the NAIP accuracy assessment. There are 45 points per water, developed, forest and agriculture class. Wetlands were not included in the NAIP accuracy assessment because the NWI is considered a “ground-truth” dataset and the NAIP wetland class was represented by a rasterized NWI.

Figure 8: Boxplot relationships between NLCD LULC classes and CCD values by waterbody size category. Smaller waterbodies had higher CCD and percent agriculture values, and lower percent forest values than medium and large waterbodies.

Figure 9: Boxplot relationships between NAIP LULC classes and CCD values by waterbody size category. Smaller waterbodies had higher CCD and percent agriculture values, and lower percent forest values than medium and large waterbodies.

Figure 10: R-squared (y-axis) and p-value (x-axis) of the 17 explanatory variables with CCD. The LULC classes are based on the NLCD, and the vertical red line indicates a 0.05 p-value. There are 13 classes with a p-value less than 0.05 (statistically significant).

Figure 11: R-squared (y-axis) and p-value (x-axis) of the 17 explanatory variables with CCD. The LULC classes are based on the NAIP, and the vertical red line indicates a 0.05 p-value. There are 12 classes with a p-value less than 0.05 (statistically significant).

Figure 12: (A) Density curve, histogram, and summary statistics of CCD 80th percentile values in the 771 waterbodies, and a (B) near 1:1-line Q-Q plot indicates distribution normality.

Figure 13: (A) NLCD MLR model residual distribution, (B) Q-Q plot, (C) Cook’s distance with a horizontal red line at 10 times the mean, and (D) diagnostics testing regression assumptions.

Figure 14: Density curve, histogram, and percent of waterbodies by risk group breakdown by NCCDI for the NLCD LULC model. The green, yellow, and red vertical lines represent
the maximum NCCDI values for the low, medium, and high-risk designations respectively (i.e. low risk are 0 ~ 0.4)

Figure 15: NAIP MLR model residual distribution, (B) Q-Q plot, (C) Cook’s distance with a horizontal red line at 10 times the mean, and (D) diagnostics testing regression assumptions...

Figure 16: Density curve, histogram, and percent of waterbodies by risk group breakdown by NCCDI for the NAIP LULC model. The green, yellow, and red vertical lines represent the maximum NCCDI values for the low, medium, and high-risk designations respectively (i.e. low risk are 0 ~ 0.4)

Figure 17: Example risk map made from NAIP LULC model MLR outputs, applied back on the 771 watersheds. The main map resolves mostly larger low risk maps, and the subset primarily shows medium and high risk in smaller watersheds

Figure 18: Density curve, histogram, and summary statistics of CCD 80th percentile values in the 36 CyanoHAB-positive Georgia Piedmont waterbodies

Figure 19: Screenshot of the cyanobacteria cell density world view dashboard. A user can select an area of interested, a cloud cover threshold, and click on water pixels to get chlorophyll-\(\alpha\), CCD, and suspended sediment concentration values.
CHAPTER 1
INTRODUCTION

Cyanobacteria, also known as blue-green algae, are single celled organisms that date back approximately 3.5 billion years when their proliferation during the Precambrian period oxygenated the early Earth’s atmosphere (Schopf, 2000). Although labeled blue-green algae, cyanobacteria are simple single-celled prokaryotic bacteria that are anatomically different than more complex, eukaryotic green-algae or “true algae.” In an episode referred to as the Great Oxidation Event, cyanobacterial activity drastically altered many Earth systems processes and allowed complex plant and animal life to emerge and evolve (Schirrmeister et. al 2015; Paerl and Paul, 2012). As autotrophic extremophiles, cyanobacteria are able to thrive in a wide diversity of ecosystems, including those in marine, freshwater, and terrestrial environments spread over great geographic and climatic ranges. Cyanobacteria can be found in almost every possible environment, ranging from benthic and lithospheric systems in the Arctic Circle to colorful hot springs in Yellowstone National Park (Schirrmeister et. al 2015; Yellowstone, 2013; Reynolds, 2006).

Cyanobacteria’s expansive evolutionary history has enhanced its survivorship by adapting a number of physical and morphological traits that allow it to dominate other aquatic competitors (Paerl and Paul, 2012). Many cyanobacterial species reach their maximum growth rate in waters above 30 degrees Celsius, while many chlorophytes, dinoflagellates, and diatoms experience rapid decreases in growth rates in waters warmer than 25 degrees Celsius (Paerl, 2014). Many species have the ability to self-regulate their buoyancy and position themselves at
an optimal location in the water column, often in the nutrient-rich warm surface waters (Paerl and Tucker, 1995; Paerl and Paul 2012). Situating into the limnetic zone benefits cyanobacteria as well because it is the ideal position to receive maximum solar radiation for photosynthetic and primary production processes. Concomitantly, cyanobacteria are shade tolerant and can photosynthesize even in low-light environments (Yamamoto and Fuh-Kwo, 2014). Cyanobacteria can survive for many years as dormant cysts to combat unfavorable extreme conditions (Potts, 1994).

Although they helped create an oxygen-rich atmosphere that has helped flora and fauna of all shapes and sizes flourish, cyanobacteria are often viewed more pejoratively in the context of the modern world. When excess nutrients, especially phosphorus and nitrogen, are introduced to a freshwater system, this eutrophic and sometime hypereutrophic situation enables cyanobacteria to grow unabated (Carlson, 1977). Freshwater eutrophication has emerged as one of the most ubiquitous environmental and social issues, and can last for considerable stretches of the year (Shi et. al, 2015). When algal populations exponentially grow and overtake a waterbody it is referred to as a Harmful Algal Blooms (HAB). If a HAB consists of cyanobacteria it is referred to as a Cyanobacteria Harmful Algal Bloom (CyanoHAB). The names are used rather interchangeably, but this study will refer to the phenomenon as a CyanoHAB.

CyanoHABs threaten human health, animal health, overall aquatic ecosystem sustainability and biodiversity, and the economic viability of affected regions (Hudnell, 2010). These negative impacts stem from the ability of many cyanobacteria genera to produce toxic compounds that have been associated with liver, kidney, digestive, skin, and neurological impairment, and even death (Carmichael, 2001; Singh et. al, 2012). Individual cyanobacteria organisms are also capable of producing more than one type of toxin, which increases risk and
complicates an already complex monitoring situation (Hudnell, 2010). Human health is also at risk when toxic cyanobacteria blooms develop in drinking water reservoirs and recreational waterbodies. Consumption of aquatic organisms containing cyanotoxins can lead to well-known illnesses such as paralytic shellfish poisoning (Hardy et. al, 2015). Humans are also at risk for dermatitis, eye irritation, and respiratory distress following exposures to cyanotoxins in recreational settings (Lin et. al, 2009). In the event of cyanotoxin aerosolization, sore throat, difficulty breathing, and worsening of preexisting respiratory conditions are possible (Hudnell, 2008). Chronic exposure to cyanotoxins can lead to long term health effects, and some cyanotoxins, such as microcystin, are suspected to possess carcinogenic characteristics (Singh et. al, 2012; Hardy et. al, 2015).

Not only do cyanotoxins affect human health, they can have detrimental effects on animal health and populations as well. Acute effects include decreased survivorship, feeding difficulties, and paralysis. Chronic effects include limited growth, diminished fecundity, biochemical alterations, and behavioral alterations (Ferrão and Kozlowsky-Suzuki, 2011). Cyanotoxins also have the potential to bioaccumulate through trophic levels. A new species of cyanobacteria, Aetokthonos hydrillicola, was shown to cause potentially fatal brain lesions in Bald Eagles after consuming American Coots (Wilde et. al, 2014). Toxicity to fauna affects the overall health of ecological systems, and can produce economic hardships for those whose livelihood depends on these faunas. There have been livestock mortality events in Georgia where farmers have had cows pass away after consumption of cyanotoxic water (Haynie et. al, 2013; Melanon, 2016). These animals have a high monetary value and the loss of even one can have serious economic implications.
Cyanobacteria proliferate most frequently when there are excessive nutrients, warm water, ample sunlight for photosynthetic activity, and quiescent water (Hudnell, 2010). These factors have strong connections to their surrounding landscapes and their individual watersheds. Watershed land management and climatology significantly dictate the role and magnitude of these factors that influence cyanobacteria growth. Changes in land use and land management associated with high nutrient loading due to agricultural and urban development are likely key factors that impair surface waters, and result in modified and simplified aquatic systems that favor cyanobacteria (Beaver et. all, 2014). Contrarily, forested and natural watersheds can provide environmental services that can uptake nutrients before they reach waterbodies.

The combination of rising temperatures and increased vertical stratification due to extended droughts “will favor cyanobacteria dominance in a wide range of aquatic ecosystems” (Hudnell, 2010). Less frequent and more intense rainfall patterns allow for periods of immense nutrient loading followed by periods of extended water stagnation. In a theoretical climate regime where winter-spring rains and summer droughts intensify, this would maximize both the nutrient loading and quiescence cyanobacteria environmental factors (Paerl and Paul, 2012). Currently, there are an average of seven CyanoHAB days per waterbody per year, and that number is expected to increase to 18-39 CyanoHAB days per year by 2090. These projections are based on a multitude of climate change scenarios, and note that the southeastern United States will experience the greatest impact in terms of costs from lost recreation (Chapra et. al, 2017). By observing the land management and climate variables of watersheds, it is possible to isolate which environmental conditions are the strongest predictors of CyanoHAB activity.

It is crucial to understand the relationship between watershed features and cyanobacteria because of their impacts to economies, human health, and ecological sustainability (Hudnell,
This research aimed to determine which landscape variables drive and trigger CyanoHABs. Geospatial and statistical techniques were used to calculate watershed-level landscape variables and identify variables that are most important to CyanoHABs. Currently, there are not any geospatial tools that predict the risk of a waterbody to CyanoHABs, and as CyanoHABs become more widespread and frequent, proactive and predictive management strategies will need to supplement reactive management strategies.

Current CyanoHAB management strategies use chemical and physical mechanisms to reduce or mitigate blooms. Copper-based algicides are one of the most commonly used methods to eradicate CyanoHABs, with chelated copper products typically being safer than traditional copper sulfate products. However, when cells die toxins existing within the cell leak into the surrounding water, creating a potentially more toxic environment. For this reason, waterbodies should continue to be closely monitored after algicides are applied. Hydrogen peroxide is also widely used, and may even be able to break down cyanotoxins, unlike copper-based methods. Other chemicals, such as Phosphlock, bind to phosphorus as they descend through the water column, dragging nutrients to the waterbody floor and significantly reduce nutrient levels without the use of toxic compounds. Those phosphorus laden sediments then must be dredged, which is an expensive and labor-intensive process. Nutrients can also be removed through the deployment of constructed and floating wetlands that harness the environmental services of wetland systems and earth materials to filter nutrients before entire a waterbody. Physically, aeration systems and fountains aim to circulate water that in theory reduces water stagnation and residence time, and therefore creates an environmental less suitable for CyanoHABs. See Hudnell, 2010 and Burtle, 2015 for a rich explanation of CyanoHAB management techniques and resources.
1.01 Analysis of Landscape Trigger Variables

Previous studies have concluded strong positive correlations between cyanobacteria blooms and warmer air temperatures (QiChao et. al, 2016; Rai and Madaiah, 2016; Hudnell, 2010), high nitrogen and phosphorus levels (Savadova, 2014; Smith 2003; Havens 2008) and decreased flushing rates due to extended drought conditions (Elliot, 2010; Paerl and Huisman, 2008; Jeppesen et. al, 2015). These scenarios represent favorable conditions for cyanobacteria proliferation based on environmental conditions necessary for cyanobacteria growth. Trends in average temperature via anthropogenic warming climate patterns are expected to continue through the mid-century mark, and will especially impact sensitive shallow waterbodies (Havens, 2016). As the warming trend continues, cyanobacteria species from tropical and subtropical areas will reach into temperate climates, disrupting food webs and altering autotrophic community composition (Savadova, 2014). The anticipated effects of climate change on temperature and precipitation patterns could play in favor of cyanobacteria’s competitive characteristics and disproportionately allow toxic CyanoHABs (Paerl and Otten, 2013). There is not a comprehensive database for cyanobacteria blooms in the United States and only a few states regularly monitor for cyanobacteria, making it difficult to study such phenomena on a historic or regional scale (Hudnell, 2010).

While many studies have conducted their research based on in-situ sampling of these variables, it is becoming increasingly documented that landscape composition is playing an important role in the transportation of nutrients into adjacent waterbodies. Agricultural and developed landscapes are commonly associated with impaired water systems, eutrophication, and loss of environmental services (Beaver et. al, 2014; Paul et. al, 2012; Katsiapi et. al, 2012). Runoff from nutrient-rich agricultural fields, lawns, and impervious urban surfaces help carry
excessive nutrients from terrestrial systems into aquatic systems. Georgia is the top producer of broilers in the United States, with over one billion birds produced each year (Georgia Farm Bureau; Ritz and William, 2013). This results in a supply of readily available nutrient-rich litter that is often applied to local pasture as a fertilizer substitute. It is a cheaper alternative to traditional fertilizers and manures, and if overly applied can results in the eutrophication and CyanoHABs in local waterbodies. Nutrient analysis of both soils and litter are available to determine effective application rates. Beaver et. al, 2014 noted that watersheds were strongly influenced by agricultural landscapes, and were more likely to produce liver-damaging hepatotoxins that are of concern for human and animal health.

The existing literature has gaps in three main areas as it relates to cyanobacteria blooms and landscape variables. Many papers included in the literature review either speak of theoretical climate change patterns (i.e. longer droughts and more intense precipitation events will encourage cyanobacteria blooms) without incorporating climate data, or have reported precipitation not directly influencing bloom dynamics (QiChao et. al, 2016; Hallstan et. al, 2013). The University of Idaho produces gridded surface meteorological data (GRIDMET) for minimum temperature, maximum temperature, and precipitation dating back to 1979. These products are updated daily and provide a rich resource for climate and weather based analyses. GRIDMET products have a spatial and temporal resolution that are well suited for researching CyanoHABs.

Secondly, the datasets used to measure landscape composition are at a scale that may not be appropriate for analyzing smaller watersheds. Landsat-based National Land Cover Dataset (NLCD) land use/land cover (LULC) maps are commonly utilized to calculate the watershed metrics for percent agriculture, forest, etc. At a 30-meter resolution and a 75-85% class accuracy
(Wickham et al., 2013), it is a convenient dataset that provides moderately high-resolution LULC maps for the entire United States (Homer et al., 2015). However, for smaller waterbodies a high-resolution map should be considered. The ability to accurately map smaller features and more complex LULC features will be vital to correctly represent the landscape composition of a watershed. The United States Department of Agriculture (USDA) National Aerial Imagery Project (NAIP) collects one-meter summer images for the United States, and their imagery is free to download. The temporal frequency of the images is dependent on state budgets, but are usually collected every one to three years. The previously red-green-blue (RGB) imagery has seen the near infrared (NIR) band added over the course of the last few years, allowing for improved classification accuracy (USDA FSA). Studies have shown 75-80% overall accuracy (Nagel and Yuan, 2016; Hayes et al., 2014) and up to 90% accuracy for finer urban features (Hayes et al., 2014; Qiu et al., 2014). Hayes et al., 2014 also points out that NLCD data is often too coarse for fine-scale research, and other high-resolution datasets are often cost-inhibitive. Although the overall accuracies of NCLD and NAIP classification data are similar, the ability of the one-meter NAIP classifications to detect finer details make it a vital tool to evaluate the land composition of watersheds, especially smaller agrarian watersheds.

Lastly, studies have predicted future risk and habitat ranges of cyanobacteria (Chen et al., 2014; Duan et al., 2017, Martinuzzi et al., 2014), but have not developed novel algorithms that incorporate readily available datasets that can be easily applied over large geographic and temporal scales. These studies provide highly detailed, site-specific information based on in-situ results, however, their results are difficult to translate to other waterbodies since they were developed with in-situ data that is not available for most waterbodies. The site-specific nature of the in-situ data is also problematic as those datasets are not always remeasured at consistent
intervals, or taken at all, which would cause issues when updating risks maps or creating risk maps for other locations. There is an absolute need for these more detailed maps, especially for drinking water reservoirs, however, the argument for this project is that there needs to be an easily adaptable CyanoHAB risk mapping method that can be created for any waterbody based on a few simple, readily available free geospatial datasets.

Addressing these literature gaps will result in a richer understanding of cyanobacteria drivers, and create a foundation for CyanoHAB risk mapping. The National Hydrography Dataset (NHD) has a layer containing most of the waterbodies for the state of Georgia. This is a nearly comprehensive dataset and includes approximately 28,000 waterbodies in the Georgia Piedmont. Due to computational restrictions that will be addressed later in this document, 771 waterbodies’ watersheds were delineated to collect 17 landscape-level variables that are hypothesized to be important to the proliferation of CyanoHABs. Then, the European Space Agency’s (ESA) Sentinel 2A satellite was used to observe the concentration of cyanobacteria cells in each waterbody. The Sentinel 2A satellite’s Multi-Spectral Imager (MSI) has the capabilities to calculate both the floating algal index (FAI) (Hu, 2009) and normalized difference chlorophyll index (NDCI) (Mishra and Mishra, 2012). Page et. al 2016 (under review) developed a cross-sensor biophysical model utilizing Sentinel 2A imagery and NDCI to create a cyanobacteria cell density (CCD) algorithm for satellite-based cyanobacteria monitoring. This method is more time and resource effective than in-situ sampling, and offers higher spatial resolution (10-20m) than other satellite-based algae monitoring methods. The introduction of these models has increased the spatial and temporal ability to monitor CyanoHABs. The 17 landscape-level variables were joined with the waterbodies’ CCD values to create a big-data
dataset that was used to perform statistical analyses to determine the most influential landscape-level drivers of CCD, and algorithms to predict waterbody CyanoHAB risk.

1.02 Google Earth Engine as a Geospatial Processing Tool

A common problem in geospatial research is the computational demands of downloading, storing, and processing large amounts of satellite imagery. Using multiple images to run simple analyses, such as spatial statistics and band indices, is computationally intensive, requiring either multiple parallel processors or waiting extended periods to receive results. This computation bottleneck is problematic for many reasons and has limited many aspects of geospatial research, restricting the spatial and temporal extends of study areas. Additionally, it limits the number of layers a study can ingest, potentially excluding important input layers.

To alleviate issues associated with data bottlenecks, the Google Earth Engine (GEE) platform was incorporated to allow exponentially faster processing time and access to a multitude of geospatial datasets in one convenient locations. GEE is a relatively new platform that has centralized multiple petabytes of geospatial datasets, including decades worth of satellite imagery and many other gridded datasets that are of interested to researchers (Gorelick et. al, 2017). GEE has both a JavaScript and a Python application programming interface (API) that allows users to produce custom code and algorithms. The code is sent from the API to Google, and the geospatial processes are completed using a farm of Google cloud computers and then returned to the API.

GEE was an integral part of this study as it made performing a NAIP-based one-meter LULC classification of the 45,976 km² Georgia Piedmont possible. A mosaicked image of the NAIP imagery for the Georgia Piedmont is about 250 gigabytes, and not even those most
powerful computers available for this study could process and classify such a large file. This was the primary use of GEE when this study began, but the exploratory spatial data analysis and big data capabilities of the platform were quickly recognized. GEE then became not just a way to classify high density aerial imagery, but a streamlined tool to incorporate several hundred study sites and 17 potential explanatory variables into the study.

Initially, this study was to include strictly LULC and climate data using ArcGIS, however, the flexibility of GEE allowed for many more variables to be ingested. This allowed a big data, exploratory spatial data approach to studying CyanoHABs. Exploratory data analysis (EDA) is a broad set of descriptive statistical tools that finds patterns in data and uses deductive reasoning, contrary to inductive reasoning found in traditional scientific method experiments, to formulate hypothesis and theories (Tukey, 1977; Oom 2013). Rather than creating a hypothesis and explicitly testing it, EDA forms hypotheses from observed patterns using inferential statistical that can then be used to evaluate anecdotal assumptions about the concept in question (Howson, 1987). The immense computational power of GEE permitted the use of 17 explanatory variables, and the opportunity to consider potentially influential datasets into the study that would have been otherwise omitted had ArcGIS been used. A list of acronyms used in this document can be found in Table 1.

1.03 Interdisciplinarity of CyanoHABs in Sustainable Food Systems

While this study focuses on the landscape-level drivers of CyanoHABs, it is important to discuss the interdisciplinarity of the broader CyanoHAB research topic and how it overall connects back to agricultural sustainability. This study was funded under the USDA Food and Agricultural Sciences National Needs Graduate and Postgraduate Fellowship (NNF) Grants
Program with the goal of creating sustainable food systems through the combined efforts of researchers. Remote sensing and GIS are one of many technologies and research approaches that contributes to the overall improvement of CyanoHAB understanding and management. Satellite sensors and statistical modeling have limitations, and while they provide an informative cog in the larger CyanoHAB research system, they need to be utilized and interpreted correctly and transparently. They should not be used unilaterally, but synergistically in combination with other tools. Risk is conceptual and bounded, and multiple data and information sources should be considered when making final CyanoHAB decisions.

Sustainability encompasses a vast number of field and sub-fields, and all types of sustainability need to be consider in the context of CyanoHABs. The most common use of sustainability is environmental suitability. The previous sections described a few of the environmental issues that lead to CyanoHABs and environmental issues that are exacerbated by CyanoHABs. Establishing further connections between CyanoHABs and watershed characteristics will help understand what needs to be done at larger geographic scales to limit future CyanoHABs. A food system cannot be considered sustainable if it deems adjacent areas inhabitable to flora and fauna, or consumes natural resources at accelerated rates.

Equally important, especially in terms of CyanoHABs and agriculture, is economic sustainability. While agricultural economics are outside the scope of this study, CyanoHABs and expedient CyanoHAB decision making is critical for the economic sustainability and longevity of many farmers. Cattle and other livestock that are exposed to CyanoHABs and have an elevated risk for effects of cyanotoxicity, both acute and chronic (Kupper et. al, 2009; Silva et. al, 2014; McGorum et. al, 2015). Profit margins in the agrarian sector can be so thin that the loss of any cattle or livestock could have negative compounding effects for a farmer and/or farming
operation. Farmers need access to up-to-date CyanoHAB information to decide whether to move cattle and livestock away from small retention ponds to safer water sources. Using alternative water sources, such a city or municipal water, is more expensive than retention ponds and require hours of labor to move the animals, reducing the number of hours spent on other farming operations. For these reasons, farmers cannot afford to keep cattle and livestock around CyanoHABs, or move them permanently to alternative water sources. Accurate information about CyanoHABs and CyanoHAB risk can help farmers best navigate the line between animal health, economic vitality, and focused labor.

We can also use livestock, particularly cattle, as sentinels to learn more about how exposures to CyanoHABs affect humans to improve public health sustainability. We do not have a clear sense of the extent of complications to human health from acute and chronic exposure to cyanotoxins. However, cattle are exposed to CyanoHABs much more regularly and could provide valuable insight. Findings from using “animal sentinels can be applied to protect people, animals and our shared environment” (Backer and Miller, 2016; Hilborn and Beasley, 2015) and build “response capacity through targeted public outreach and prevention activities” (Backer et al., 2015).

Finally, without appropriate outreach efforts, any new tools or decision aiding information will not get to where it needs to go to be impactful. There needs to be increased awareness about tools as they become available, as well as robust, simplified translations of their products and outputs. If tools are complicated or require intense interpretation, they will lose effectiveness and ultimately not be used as often as a more simplified counterpart. Education and outreach is similarly important as tool research and development itself, and this study was crafted and executed with that objective in mind. The modern food system is expansive and
infinitely nuanced, and it is plain to see the need for interdisciplinary studies and outreach efforts to truly make a difference in CyanoHAB management and sustainable agriculture and food systems.
Table 1: List of acronyms used in this document.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAB</td>
<td>Harmful Algal Bloom</td>
</tr>
<tr>
<td>CyanoHAB</td>
<td>Cyanobacteria Harmful Algal Bloom</td>
</tr>
<tr>
<td>GRIDMET</td>
<td>Gridded Surface Meteorological Data</td>
</tr>
<tr>
<td>NLCD</td>
<td>National Land Cover Dataset</td>
</tr>
<tr>
<td>LULC</td>
<td>Land Use Land Cover</td>
</tr>
<tr>
<td>USDA</td>
<td>United States Department of Agriculture</td>
</tr>
<tr>
<td>NAIP</td>
<td>National Aerial Imagery Program</td>
</tr>
<tr>
<td>RGB</td>
<td>Red-Green-Blue</td>
</tr>
<tr>
<td>NIR</td>
<td>Near Infrared</td>
</tr>
<tr>
<td>NHD</td>
<td>National Hydrography Dataset</td>
</tr>
<tr>
<td>ESA</td>
<td>European Space Agency</td>
</tr>
<tr>
<td>MSI</td>
<td>Multispectral Instrument</td>
</tr>
<tr>
<td>FAI</td>
<td>Floating Algae Index</td>
</tr>
<tr>
<td>NDCI</td>
<td>Normalized Difference Chlorophyll Index</td>
</tr>
<tr>
<td>CCD</td>
<td>Cyanobacteria Cell Density</td>
</tr>
<tr>
<td>GEE</td>
<td>Google Earth Engine</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>EDA</td>
<td>Exploratory Data Analysis</td>
</tr>
<tr>
<td>SWIR</td>
<td>Shortwave Infrared</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>DEM</td>
<td>Digital Elevation Model</td>
</tr>
<tr>
<td>NWI</td>
<td>National Waterbody Inventory</td>
</tr>
<tr>
<td>NDVI</td>
<td>Normalized Difference Vegetation Index</td>
</tr>
<tr>
<td>NDWI</td>
<td>Normalized Difference Water Index</td>
</tr>
<tr>
<td>PRISM</td>
<td>Parameter-Elevation Relationships and Independent Slope Model</td>
</tr>
<tr>
<td>EVI</td>
<td>Enhanced Vegetation Index</td>
</tr>
<tr>
<td>NED</td>
<td>National Elevation Dataset</td>
</tr>
<tr>
<td>gSSURGO</td>
<td>Gridded Soil Survey Geographic Database</td>
</tr>
<tr>
<td>SSURGO</td>
<td>Soil Survey Geographic Database</td>
</tr>
<tr>
<td>NRCS</td>
<td>Natural Resources Conservation Services</td>
</tr>
<tr>
<td>USGS</td>
<td>United States Geological Survey</td>
</tr>
<tr>
<td>SPARROW</td>
<td>Spatially Referenced Regressions on Watershed Attributes</td>
</tr>
<tr>
<td>HUC</td>
<td>Hydrologic Unit Code</td>
</tr>
<tr>
<td>MLR</td>
<td>Multiple Linear Regression</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>NCCDI</td>
<td>Normalized Cyanobacteria Cell Density Index</td>
</tr>
<tr>
<td>SSC</td>
<td>Suspended Sediment Concentration</td>
</tr>
<tr>
<td>MODIS</td>
<td>Moderate-Resolution Imaging Spectroradiometer</td>
</tr>
<tr>
<td>MERIS</td>
<td>Medium Resolution Imaging Spectrometer</td>
</tr>
</tbody>
</table>
CHAPTER 2
MATERIALS AND METHODS

2.01 Study Area

The Piedmont is one of five physiographic regions in Georgia, along with the Coastal Plain, Blue Ridge, Ridge and Valley, and the Plateau. Although it is the second largest region by area, it has the most inhabitants and is situated between the Appalachian Mountains to the north and the Coastal Plain, demarcated by the Atlantic Seaboard Fall Line, to the south (Jackson and Stakes, 2004). The Georgia Piedmont has productive soils that are ideal of agriculture. Large-scale cotton, wheat, and soybean farming can be found through the region, along with a healthy beef and dairy cattle industry. Especially important to the region’s economy is the chicken broiler industry, which has earned the nickname of the “Poultry Capital of the World” (Jackson and Stakes, 2004). Poor agricultural practices during the 20th century caused severe erosion of the Piedmont’s soils, and a loss of nutrients and vital topsoil (Sutter, 2010). To mitigate this loss, farmers are forced to apply more synthetic fertilizers, and more commonly, cheap, widely available, nutrient-rich broiler litter to ensure economically viable crop yields. Surface runoff after large precipitation events cause nutrients from the synthetic fertilizers and broiler litter to be transported downslope into water catchments. Generally warm temperatures, intense fertilizer application, and a rapidly urbanizing and populating landscape make the Georgia Piedmont highly susceptible to CyanoHABs.

A stratified random sample was taken from the approximately 28,000 waterbodies in the NHD dataset (USGS et. al, 2017), divided into three size categories based on their area: small (<
0.01 km\(^2\), medium (\(> 0.01 \text{ km}^2, < 0.1 \text{ km}^2\)), and large (\(> 0.1 \text{ km}^2, < 1 \text{ km}^2\)). Waterbody size was limited to one square kilometer because the watersheds of those waterbodies were expansive, often encompassing large fractions of the Georgia Piedmont and being composed of many sub-watersheds. Originally 700 waterbodies were selected from each category for a total of 2,100, however, it was later discovered that not all the NHD waterbodies were permanent surface water, resulting in the omission of 1,329 waterbodies. These omitted waterbodies were often semi-permanent and marshy wetlands. Waterbodies were partitioned as surface waters or non-surface waters based on a shortwave infrared (SWIR) threshold. This left a total of 771 waterbodies, and new category samples sizes of 346, 257, and 168 for small, medium, and large respectively (Figure 1). The categories were created to ensure a wide variety of waterbody sizes in the study. The three categories were combined into one dataset for all further analyses (Figure 2) because of low adjusted R-squared values (< 0.10) for size-specific CyanoHAB risk models (see sections 2.06 and 3.02 for description of modeling process and results). The outline of the Georgia Piedmont was digitized from the “Geographic Regions of Georgia” map provided by GeorgiaInfo [http://georgiainfo.galileo.usg.edu/topics/geography/article/geographic-regions-of-georgia].

2.02 Watershed Delineation Automation

A watershed is a basic hydrological unit that represents the common area where all linear water features and precipitation collect (Perlman, 2016). The watershed is an important area to accurately define for the 771 waterbodies in this study, as they bound the landscape-level data that was collected. There are many ways to delineate watersheds using Geographic Information
Figure 1: Colored coded map of the 771 waterbodies used in this study by size category. The Georgia Piedmont is outlined by a thick black line.
Figure 2: Single color map of the 771 waterbodies used in this study. The Georgia Piedmont is outlined by a thick black line. The background image is a Sentinel 2A satellite image from July 26th, 2016. This is the image that was used to calculate CCD for the waterbodies, which is why waterbodies west of the image in the Georgia Piedmont were not included in this study.
Systems (GIS), including TOPODATA drainage networks (Mantelli et. al, 2011), the Multi-Watershed Delineation Tool from Utah State University, the ArcHydro extension of ArcGIS (Konadu and Fosu 2010), and the Watershed tool within the standard ArcGIS Hydrology toolbox. These tools require either ancillary datasets or customized workspace environments to run properly, which complicates the process, especially when different, incompatible versions of software or datasets are being used. Additionally, they can take large amounts of manual intervention and data manipulation to perform just one delineation. Automatic watershed delineation has been performed in the past (Mantelli et. al, 2011; Baker et. al, 2006), but there is not a publicly available tool that is considered the “standard tool.” To delineate the 771 watersheds for this study, an automated tool was created in ArcGIS Model Builder that iterates the Watershed tool workflow in the Hydrology toolbox. Figure 3 shows a workflow diagram for the methodology.

To delineate a watershed, there are two necessary data layers; a flow accumulation raster and a pour point. The flow accumulation raster is created from the digital elevation model (DEM)-based flow direction raster, and represents the number of pixels that cumulatively flow into each pixel. It creates a dendritic pattern akin to a stream network, with larger rivers having higher flow accumulation values than creeks, having higher values than flat land features. The pour point indicates the location at which the watershed delineation is desired. This point is defined by the highest flow accumulation value within a waterbody, and is used as a starting point to trace pixels upstream to find which pixels flow into it. Pixels that flow into the pour point are then considered to be a part of its watershed. Using the ArcGIS Hydrology toolbox is a highly manual process. To streamline this process, two ModelBuilder tools were created; one to automatically create pour points at the highest flow accumulation pixel location, and one to
Figure 3: Workflow for watershed delineation automation. This process was carried out using ArcGIS ModelBuilder.
iterate through those pour points and delineate each watershed. The two models were used synergistically following the workflow in Figure 3 to delineate the watersheds for this study’s 771 waterbodies.

2.03 NAIP LULC Classification

The NLCD is a nation-wide pixel-based classification based on a combination of cloud-free images from the Landsat satellite series, and currently exists for 1992, 2001, 2006, and 2011, with the 2016 dataset is slated to be released in the near future. It follows the Anderson level II classification scheme and has 20 classes, 15 of which can be found in the Georgia Piedmont (USGS, 2017). A level I classification include broad LULC categories, such as forest, agriculture, and developed. A level II classification scheme subdivides those larger classes into more specific LULC types, such as evergreen forest, soybean agriculture, and medium-density developed. The NLCD is commonly used to study LULC as it is freely available, thoroughly assessed, and its 30-meter resolution is an appropriate scale for most studies. While the NLCD has an overall level I accuracy of 70-91% (Stehman et. al, 2003; Wickham et. al, 2010; Wickham et. al 2013; Wickham et. al, 2017), 30-meter pixels may be too large to efficiently capture to LULC characteristics of smaller watersheds with highly heterogeneous matrices.

NAIP imagery is somewhat limited in its ability to resolve a level II classification scheme because of its four-band architecture that does not capture class-specific spectral variability. For this reason and the purpose of creating a general relationship between CyanoHABs and broad LULC characteristics, a level I classification scheme was used for the NAIP LULC classification. The classes followed a simplified version of the Anderson I classification scheme,
and included water, developed, forest, agriculture, and wetland. See Table 2 for a translation of level II to level I classes. The same level I classification scheme was applied to the 2011 NLCD for parallel analyses for overall assessment and relationship with CyanoHABs. The 2011 NLCD was the most recent dataset available at the time of the study.

Class-exhaustive NAIP based LULC classifications have been made using maximum likelihood, random forest, and support vector machine classification algorithms in the past (Hayes et. al, 2014; Maxwell et. al, 2014). GEE currently supports all three of these classification algorithms with customizable parameterization and settings for all. This study originally tested these three classification algorithms on a clipped mosaic image of 2015 NAIP imagery for the Georgia Piedmont, and preliminary visual assessments using these methods were not satisfactory. Additionally, using the classification algorithms takes an immense amount of computational power, and even GEE failed at times to create final classification maps. To create as accurate of a map as possible, ancillary vector data were merged with the NAIP imagery, and a user-defined spectral feature space was created to perform classification based on single-band and band ratio thresholding. Using a series of thresholds also helped to alleviate computational overloading associated with the classification algorithms. Wetland data were downloaded from the National Wetland Inventory (NWI), surface water data were downloaded from the NHD, and roads data were downloaded from the Census Bureau TIGER/Line shapefile web interface.

All of these datasets are vectors, and they were rasterized at a one-meter resolution and uploaded to GEE. Normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) bands were calculated from Rouse et. al, 1973 and McFeeters, 1996 respectively, and composited with the blue, green, red, and near infrared NAIP bands. The user-defined spectral feature space was used these rulesets to create a class-exhaustive level I
Table 2: Grouping of level II classification classes into level I classification classes.

<table>
<thead>
<tr>
<th>Level I Class</th>
<th>Level II Classes Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>Open Water</td>
</tr>
<tr>
<td>Developed</td>
<td>Barren Land; Open Space, Low Intensity, Medium Intensity, High Intensity developed</td>
</tr>
<tr>
<td>Forest</td>
<td>Deciduous, Evergreen, Mixed Forest, Shrub/Scrub</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Grassland/Herbaceous, Pasture/Hay, Cultivated Crops</td>
</tr>
<tr>
<td>Wetland</td>
<td>Woody Wetlands, Emergent Herbaceous Wetlands</td>
</tr>
</tbody>
</table>
classification, and then the NWI, NHD, and Census roads rasters were burned in on top of the classification.

2.04 Google Earth Engine Processing

Although GEE has the computational power to classify the 250 gigabyte Georgia Piedmont, GEE encountered issues when classifying the imagery and then using it to calculate the percent LULC for the 771 watersheds. The NAIP classification was exported as a GEE asset, which is the equivalent of having a “local” file. Rather than reclassifying the NAIP imagery every time the code is run, exporting as a GEE asset allows GEE to pull the classified image from memory. Additionally, the 771 watersheds contained a high volume of geometric points that compose the polygon shapefiles. This caused GEE to crash, and to ease the explanatory variable calculation, the watershed shapefiles were simplified. A simplify function exists within GEE that reduces the number of polygon vertices, and this function allowed GEE to complete the watershed calculation of the 17 explanatory variables.

2.05 Description of Google Earth Engine Datasets

The 17 CCD explanatory variables came from a variety of sources, both included and excluded in the preexisting GEE datasets. GEE houses a diverse collection of geospatial datasets, not just satellite imagery, and users can request new datasets for GEE to upload. Any dataset used in this analysis that was not already included in the GEE repositories was uploaded as an asset, then incorporated into the GEE code. An overview of the 17 explanatory variables, their source and quantifying statistic are available in Table 3, and the data sources that have not been
Table 3: List of the explanatory variables used in the CCD analysis. All variables are measured at the watershed level, except for waterbody area.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Source</th>
<th>Watershed Statistic</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterbody Area</td>
<td>NHD</td>
<td>N/A</td>
<td>Square Km</td>
</tr>
<tr>
<td>Percent Agriculture</td>
<td>NLCD 2011/NAIP 2015</td>
<td>Ratio</td>
<td>Unitless</td>
</tr>
<tr>
<td>Percent Developed</td>
<td>NLCD 2011/NAIP 2015</td>
<td>Ratio</td>
<td>Unitless</td>
</tr>
<tr>
<td>Percent Forest</td>
<td>NLCD 2011/NAIP 2015</td>
<td>Ratio</td>
<td>Unitless</td>
</tr>
<tr>
<td>Dry Season Minimum Temp</td>
<td>GRIDMET, Jan – March, 2016</td>
<td>Mean</td>
<td>Celsius</td>
</tr>
<tr>
<td>Dry Season Maximum Temp</td>
<td>GRIDMET, Jan – March, 2016</td>
<td>Mean</td>
<td>Celsius</td>
</tr>
<tr>
<td>Dry Season Precipitation</td>
<td>GRIDMET, Jan – March, 2016</td>
<td>Sum</td>
<td>Millimeter</td>
</tr>
<tr>
<td>Wet Season Minimum Temp</td>
<td>GRIDMET, April – July 26, 2016</td>
<td>Mean</td>
<td>Celsius</td>
</tr>
<tr>
<td>Wet Season Maximum Temp</td>
<td>GRIDMET, April – July 26, 2016</td>
<td>Mean</td>
<td>Celsius</td>
</tr>
<tr>
<td>Wet Season Precipitation</td>
<td>GRIDMET, April – July 26, 2016</td>
<td>Sum</td>
<td>Millimeter</td>
</tr>
<tr>
<td>EVI Dry Season EVI Anomaly</td>
<td>Aqua MODIS, Jan – March, 2016</td>
<td>Difference</td>
<td>Unitless</td>
</tr>
<tr>
<td>Slope</td>
<td>National Elevation Dataset</td>
<td>Mean</td>
<td>Percent</td>
</tr>
<tr>
<td>Aspect</td>
<td>National Elevation Dataset</td>
<td>Mean</td>
<td>Degrees clockwise from N</td>
</tr>
<tr>
<td>Percent Clay Soil</td>
<td>gSSURGO</td>
<td>Mean</td>
<td>Kilograms</td>
</tr>
<tr>
<td>Percent Nitrogen</td>
<td>SPARROW</td>
<td>Mean</td>
<td>Kilograms</td>
</tr>
<tr>
<td>Percent Phosphorus</td>
<td>SPARROW</td>
<td>Mean</td>
<td>Kilograms</td>
</tr>
<tr>
<td>Nitrogen/Phosphorus Ratio</td>
<td>SPARROW</td>
<td>Ratio</td>
<td>Unitless</td>
</tr>
</tbody>
</table>
already discussed will be briefly introduced below. The waterbody area was not calculated using GEE, it is a preexisting metadata property within the NHD dataset.

### 2.05.1 GRIDMET

All climate and weather variables were processed from GRIDMET datasets. GRIDMET is available through GEE and combines Parameter-elevation Relationships and Independent Slope Model (PRISM) data with daily gauge-based precipitation observations to create a gridded meteorological dataset that have both high temporal (daily) and spatial (~4 kilometers) resolution (Abatzoglou, 2013). GRIDMET products cover the conterminous United States and includes daily temperature minimum, temperature maximum, total precipitation, humidity, 10-meter wind velocity, and surface downward shortwave radiation. These products are available from 1979 – present day and are a rich source of climate and weather variables.

The average temperature minimum, temperature maximum, and total precipitation variables were divided into two groups; winter season and summer season. The reasoning for this is the aforementioned theoretical climate paradigm where increased precipitation and warmer temperatures in the winter, and decreased precipitation and extreme temperatures in the summer would favor cyanobacteria growth. Splitting the climate and weather data into winter (January 1, 2016 – March 31, 2016) and summer (April 1, 2016 – July 26, 2016, date of Sentinel 2A image) allows a comparison of the antecedent climate conditions by time range. April 1\textsuperscript{st} was chosen as the cutoff between the winter and summer season based on the climatology of the Georgia Piedmont in Figure 4; a 35-year climatology of GRIDMET daily high temperature data. In April, high temperatures start to reach 25 degrees Celsius, the temperature at which chlorophyte production decreases and cyanobacteria production increases (Paerl, 2014).
Figure 4: Climatology histogram for the Georgia Piedmont based on GRIDMET daily high temperature data from 1980 – 2015. The daily images for each month of each year during the 35-year period were averaged to make one average maximum temperature raster per month, and like-months were averaged to create their respective histogram.
2.05.2 Aqua MODIS 16-Day EVI Composites

The Moderate-Resolution Imaging Spectroradiometer (MODIS) MYD13Q1 product was used to create an enhanced vegetation index (EVI) anomaly map for the 2016 winter season (January 1 – March 31, 2016). MYD13Q is a 16-day composite product that maps vegetation greenness using multiple vegetative indices at 250-meter spatial resolutions (NASA, 2017). The purpose of this layer is to quantify abnormally high EVI values during the winter, in which healthy green vegetation should be nonexistent. An abnormally high value would indicate an area with excessive nutrients that are allowing plants to green up earlier than their natural phenological cycle. Areas with excessive nutrients could represent a watershed that would be prone to CyanoHABs after a large rain event. To establish a phenological baseline, MYD13Q EVI images from January 1 – March 31 during 2003-2015 were averaged by year, and then averaged overall. To calculate the EVI anomaly, the average MYD13Q value from January 1 – March 31, 2016 were subtracted from the average 2003-2015 EVI baseline. The MYD13Q library is available in GEE. EVI usually has a -1 to 1 range, however, this product is scaled from 0 to 10,000 in GEE.

2.05.3 National Elevation Dataset

The National Elevation Dataset (NED) is an elevation product for the United States and its territories at a variety of resolutions. In the conterminous United States, 30-meter and 10-meter DEMs are available, and 3-meter resolution DEMs are sparingly available for areas associated with high coastal risks (USGS, 2015). The 10-meter NED DEM is available in GEE, and GEE has specialized algorithms to calculate slope and aspect from a DEM. Slope (percent, 0-90) was calculated to represent the steepness of a watershed and its potential for excessive
runoff, and aspect was calculated to represent the exposure of soils to sunlight that could affect soil moisture and percolating capacity.

2.05.4 gSSURGO

The gridded soil survey geographic database (gSSURGO) offers a raster version of the traditional, vector-based soil survey geographic database (SSURGO). SSURGO is a regional, state, and national database that provides physical and chemical characteristics of soils through surveying and laboratory efforts (USDA NRCS, 2015). This study was interested in the percent of clay soils in waterbody watersheds because of clay soil’s ability to bind with phosphorus and nitrogen and transport the nutrients into surface water (Grift et. al, 2016; Van der Salm et. al, 2012). gSSURGO data was not available in GEE, so it was downloaded from the USDA Natural Resource Conservation Services (NRCS) Soils website. The dataset is a combination of many lookup tables and geodatabases, and is generally difficult to navigate. The USDA NRCS created a Soil Data Development Tools add-on for ArcGIS that allows users to click drops down menus to pull soil data maps, like percent clay soil, instead of dealing with complicated geodatabases. This tool was used to create a 10-meter percent clay map for the Georgia Piedmont, and the was uploaded to GEE as an asset.

2.05.5 SPARROW

The United States Geological Survey (USGS) spatially referenced regressions on watershed attributes (SPARROW) models were developed to track the movement of nitrogen and phosphorus from source locations into surface waters to explain spatial patterns of water quality in relation to anthropogenic activities and environmental processes (USGS NAWQA,
SPARROW is different than most other datasets used in this study because it is created at a hydrologic unit code (HUC) 12 scale, a watershed-shaped shapefile. Although SPARROW is a semi-homogeneous vector dataset, it provides yearly estimation for total yield of nitrogen and phosphorus (kg), which could help to explain spatial patterns of surface water eutrophication. Additionally, nitrogen to phosphorus ratios can differentiate eutrophic conditions that may favor chlorophytes (high ratio) or cyanobacteria (low ratios) due to cyanobacteria’s ability to fix their own nitrogen (Chislock et al. 2014). There are 672 HUC-12s in the Georgia Piedmont, which were rasterized at 10-meter resolution to provide comparable spatial resolution while limiting file size. Layers were created for total nitrogen, total phosphorus, and the ratio of total nitrogen to total phosphorus, and then uploaded to GEE to as an asset.

2.05.6 Cyanobacteria Cell Density using Sentinel 2A

The previous sections described the processes of collecting watershed-level data for the study sites, and to relate that landscape data to CyanoHABs, satellite imagery and remote sensing techniques were used to calculate CCD for each of the 771 waterbodies. ESA’s new Sentinel 2A MSI sensor has the ideal spatial (10-20m) and spectral (narrow bands at 665 nm and 705 nm) resolutions to monitor CyanoHABs and quantify CCD. The least cloudy Sentinel 2A image from summer 2016 was taken on July 26, 2016, and was corrected for ozone and Rayleigh scattering. The result was a six-band Rayleigh corrected reflectance image based on methods from Page, 2016, Mishra et al., 2005, and Mishra et al., 2014. The six bands were the blue, green, red, near infrared, and two SWIR bands. To mask non-water pixels, the normalized difference water index (NDWI) was calculated from McFeeters, 1996 using Equation (1).
\[ \text{NDWI} = \frac{(\text{Green} - \text{NIR})}{(\text{Green} + \text{NIR})} \tag{1} \]

where “Green” is the green band (band 3) and “NIR” is the near infrared band (band 8) of the Rayleigh corrected image. Water was defined as pixels that had NDWI values greater than -0.45, blue (band 2) values less than 0.02, green (band 3) values less than 0.035, and SWIR (band 11) values less than 0.05.

The following processes were then performed on the Rayleigh corrected reflectance image. Equation (2) was used to calculate the NDCI from the corrected Sentinel 2A bands based on an adapted method from Mishra and Mishra 2012. The original NDCI equation from Mishra and Mishra 2012 used wavelengths that Sentinel 2A does not have, so the Sentinel bands closest to those wavelengths were substituted.

\[ \text{NDCI}_{\text{Sen2A}} \propto \frac{[\rho_{\text{rc}}(705) - \rho_{\text{rc}}(665)]}{[\rho_{\text{rc}}(705) + \rho_{\text{rc}}(665)]} \tag{2} \]

where \( \rho_{\text{rc}}(\lambda) \) is the Rayleigh corrected reflectance at the given wavelength in nanometers.

Equation (3) uses the previously calculated Sentinel 2A NDCI values to convert the index values into per-pixel chlorophyll-a concentrations.

\[ \text{Chl-}a = 14.039 + 86.115 * \text{NDCI}_{\text{Sen2A}} + 194.325 * (\text{NDCI}_{\text{Sen2A}})^2 \tag{3} \]

Finally, Equation (4) (Page et. al, 2017) was used to calculate the CCD from the Sentinel 2A derived chlorophyll-a image. The resulting image displayed the CCD for water pixels in the Georgia Piedmont (see Figure 19 for a colorized subset of a CCD image).
\[
\text{CCD}_{\text{Sen2A}} \text{ (cells / mL)} = 4,989.5 \times \text{Chl}-d_{\text{Sen2A}} - 131,742 \ (4)
\]

In an effort to appropriately quantify and represent the CCD for a waterbody, the value of the 80th percentile pixel was used. This number was agreed upon because it was believed that using the maximum CCD value could be misleading based on erroneous pixel values or misclassified water pixels, and the mean or median could fail to capture high CCD values during a bloom. The 80th percentile was a compromise between the two metrics as a way to reduce false maxima while still representing CyanoHAB conditions.

2.06 Cyanobacteria Cell Density Regression

GEE was used to calculate waterbody CCD and the 17 explanatory watershed variables, and an example of the table format can be seen in Table 4. This data represents that actual data collected for the seven waterbodies with the highest CCD values. The 17 explanatory variables were logically selected based on their suspected importance to CyanoHAB from the literature. Multiple linear regressions (MLR) were used to quantify the relationship between the 17 landscape variables and the CCD values. To create a simple risk map for CyanoHABs, this study aimed to have the strongest regression strength with the fewest number of variables, commonly referred to as a parsimonious model (Richards et. al, 2011). With 17 explanatory variables, there are thousands of variable combinations available. There is a “dredge” function within the “MuMln” R package that performs a MLR on all combination of variables, and then outputs the models by their Akaike information criterion (AIC). AIC is a statistical metric that represents model confidence relative to the other combinations (Richards et. al, 2011). The variable combination with the lowest AIC score is considered the best model. This process was
Table 4: Example table of the seven highest CCD values (total study n = 771).

<table>
<thead>
<tr>
<th>ID</th>
<th>CCD</th>
<th>Area</th>
<th>Slope</th>
<th>Aspect</th>
<th>Sum Min</th>
<th>Sum Max</th>
<th>Sum Precip</th>
<th>Winter Min</th>
<th>Winter Max</th>
<th>Winter Precip</th>
<th>% Ag</th>
<th>% Forest</th>
<th>% Developed</th>
<th>% Clay</th>
<th>Nitrogen</th>
<th>Phosphorus</th>
<th>N/P Ratio</th>
<th>EVI Anomaly</th>
</tr>
</thead>
<tbody>
<tr>
<td>55250581</td>
<td>4065508.72</td>
<td>0.21</td>
<td>2.20</td>
<td>175.52</td>
<td>17.18</td>
<td>29.80</td>
<td>208.75</td>
<td>3.78</td>
<td>16.77</td>
<td>265.55</td>
<td>10.39</td>
<td>53.67</td>
<td>27.89</td>
<td>36.68</td>
<td>286.85</td>
<td>43.82</td>
<td>6.55</td>
<td>161.32</td>
</tr>
<tr>
<td>55249449</td>
<td>3681365.67</td>
<td>0.06</td>
<td>2.85</td>
<td>191.90</td>
<td>16.71</td>
<td>29.81</td>
<td>219.84</td>
<td>3.69</td>
<td>16.33</td>
<td>300.22</td>
<td>19.29</td>
<td>52.06</td>
<td>23.04</td>
<td>32.90</td>
<td>785.56</td>
<td>68.95</td>
<td>11.39</td>
<td>331.79</td>
</tr>
<tr>
<td>49692793</td>
<td>3538048.46</td>
<td>0.13</td>
<td>3.40</td>
<td>208.30</td>
<td>16.63</td>
<td>30.27</td>
<td>262.05</td>
<td>2.76</td>
<td>17.16</td>
<td>254.80</td>
<td>20.28</td>
<td>60.43</td>
<td>14.86</td>
<td>34.01</td>
<td>192.64</td>
<td>24.82</td>
<td>7.76</td>
<td>387.87</td>
</tr>
<tr>
<td>55255855</td>
<td>342484.52</td>
<td>0.01</td>
<td>2.89</td>
<td>235.41</td>
<td>17.66</td>
<td>30.99</td>
<td>315.80</td>
<td>5.06</td>
<td>18.52</td>
<td>188.20</td>
<td>76.24</td>
<td>12.43</td>
<td>7.05</td>
<td>21.50</td>
<td>190.57</td>
<td>16.19</td>
<td>11.77</td>
<td>438.66</td>
</tr>
<tr>
<td>52042747</td>
<td>341493.02</td>
<td>0.01</td>
<td>2.80</td>
<td>294.62</td>
<td>16.52</td>
<td>30.29</td>
<td>272.24</td>
<td>2.43</td>
<td>17.36</td>
<td>248.79</td>
<td>45.20</td>
<td>35.33</td>
<td>9.98</td>
<td>37.70</td>
<td>407.08</td>
<td>37.67</td>
<td>10.81</td>
<td>2.29</td>
</tr>
<tr>
<td>52036498</td>
<td>336604.15</td>
<td>0.01</td>
<td>0.85</td>
<td>117.09</td>
<td>16.71</td>
<td>29.74</td>
<td>261.79</td>
<td>3.15</td>
<td>15.74</td>
<td>265.22</td>
<td>13.02</td>
<td>69.29</td>
<td>11.83</td>
<td>27.97</td>
<td>359.18</td>
<td>47.98</td>
<td>7.49</td>
<td>51.20</td>
</tr>
<tr>
<td>41276112</td>
<td>310947.86</td>
<td>0.01</td>
<td>2.04</td>
<td>117.76</td>
<td>15.70</td>
<td>28.15</td>
<td>270.75</td>
<td>2.24</td>
<td>13.82</td>
<td>294.68</td>
<td>55.67</td>
<td>14.03</td>
<td>12.13</td>
<td>32.39</td>
<td>486.23</td>
<td>75.86</td>
<td>6.41</td>
<td>658.25</td>
</tr>
</tbody>
</table>
run twice, once for CCD and the 17 explanatory variables including NAIP-based LULC percentages, and once for CCD and the 17 explanatory variables including the NLCD-based LULC percentages.

MLRs were then run for the NAIP and NLCD datasets and their respective best model variables. The MLR outputs gave the adjusted-$R^2$ value and the individual variable p-values. At that point, the models included 9 and 8 variables for NAIP and NLCD respectively, so to reduce the number of variables in the models any variables with a p-value above 0.05 were removed. This left both models with five variables each. Outliers were removed from the datasets based on the Cook’s distance of the model residuals, and any observation whose Cook’s distance was greater than 10 times that of the overall mean were removed. Cook’s distance represents the magnitude fitted values would change if that point were omitted, and is one method used to detect outliers (Wang et al., 2017). A last MLR for the NAIP and NLCD datasets were run with the final five variables and removed outliers. The models were then tested for five assumptions of linear regressions; normal dependent variables, normal model residuals, no multicollinearity, homogeneous variance of residuals, and no outliers.

Figure 5 shows the locations of 36 waterbodies that have had in-situ confirmed CyanoHABs. The 36 locations were provided by Dr. Susan Wilde of the University of Georgia’s Warnell School of Natural Resources and Forestry. Since they are locations of known CyanoHAB activity, their CCD values were also calculated using Equation 4. Their values were 1) used to validate Equation 4 by showing non-zero CCD values, 2) to see if elevated CCD values were detected in known CyanoHAB locations, and 3) compare the distribution of CCD values in known CyanoHAB locations to those of the 771 waterbodies. A total of 59 CyanoHAB
Figure 5: Cyanobacteria-positive locations in the Georgia Piedmont confirmed by *in-situ* sampling by Dr. Susan Wilde from the University of Georgia.
positive waterbodies have been identified across Georgia, however 36 fell within the July 26th, 2016 Sentinel 2A image swath that was used to calculate CCD.

2.07 Cyanobacteria Cell Density Risk Assignment

The variable coefficient outputs of the final NAIP and NLCD MLR models were compiled to create an equation to predict the CCD in a waterbody based on five landscape variables. That equation was then applied back to the original data table to get a predicted CCD value for each waterbody. A normalized CCD 80th percentile index (NCCDI) was created by dividing each waterbody’s CCD value by the highest waterbody CCD value, resulting in a 0 to 1 scale. A Jenk’s natural break decided the breakoff points for three NCCDI classes, which translated to “low,” “medium,” and “high” risk. This process could be applied to other delineated waterbodies across the greater southeastern United States as well.
CHAPTER 3
RESULTS AND DISCUSSION

3.01 NAIP and NLCD Accuracy Assessment

Subset examples of the NAIP and NLCD LULC classifications can be seen in Figure 6. Visually, it is apparent the level of detail present in the NAIP classification compared to the NLCD classification. Smaller features such as roads and individual trees were resolved in the NAIP classification, whereas, in NLCD, features are clumped together and finer details are lost.

Using GEE, 45 random points were assigned per class for an accuracy assessment. Figure 7 shows the points used for the NAIP accuracy assessment. Using the default Google Earth satellite basemap, “ground-truth” LULC classes were assigned to the points. Consumer’s, producer’s, and overall accuracy, and Cohen’s kappa coefficient were calculated for the NAIP and NLCD classifications using accuracy assessment and confusion matrices functions in GEE. Consumer’s accuracy is the probability that a pixel classified on the image represents that category on the ground. Producer’s accuracy is the probability of a reference pixel being correctly classified. Overall accuracy is the ratio to correctly classified pixels to total pixels. Cohen’s kappa coefficient is the actual accuracy of the map compared to the expected accuracy of the map. Table 5 highlights the accuracy and kappa outputs.

The NAIP classification had higher consumer’s accuracy for all comparable classes, and a higher overall accuracy and kappa coefficient. The NAIP classification had an overall accuracy...
Figure 6: Visual comparison of level I 30-meter NLCD (A, C) and one-meter NAIP (B, D) classifications. A and B are from the same extent, as are C and D. Water is blue, developed is red, forest is green, agriculture is yellow, and wetland is cyan. The images are semi-transparent to allow satellite imagery to be visible beneath.
Figure 7: Randomly generated points that were used for the NAIP accuracy assessment. There are 45 points per water, developed, forest and agriculture class. Wetlands were not included in the NAIP accuracy assessment because the NWI is considered a “ground-truth” dataset and the NAIP wetland class was represented by a rasterized NWI.
Table 5: Classification accuracy breakdown for the NLCD (A) and NAIP (B) LULC maps. All NAIP LULC classes had a higher consumer’s accuracy than their respective NLCD counterparts, and overall the NAIP was 19.27% more accurate.

<table>
<thead>
<tr>
<th>A) NLCD - Land Cover Type</th>
<th>Consumer’s Accuracy</th>
<th>Producer’s Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>80.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Developed</td>
<td>66.67%</td>
<td>71.43%</td>
</tr>
<tr>
<td>Forest</td>
<td>93.33%</td>
<td>43.75%</td>
</tr>
<tr>
<td>Agriculture</td>
<td>37.78%</td>
<td>77.27%</td>
</tr>
<tr>
<td>Wetland</td>
<td>64.44%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

**Overall Accuracy – 68.44%**

**Overall Kappa Coefficient – 0.605**

<table>
<thead>
<tr>
<th>B) NAIP - Land Cover Type</th>
<th>Consumer’s Accuracy</th>
<th>Producer’s Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>84.44%</td>
<td>97.44%</td>
</tr>
<tr>
<td>Developed</td>
<td>84.78%</td>
<td>90.70%</td>
</tr>
<tr>
<td>Forest</td>
<td>93.48%</td>
<td>75.44%</td>
</tr>
<tr>
<td>Agriculture</td>
<td>92.50%</td>
<td>94.87%</td>
</tr>
</tbody>
</table>

**Overall Accuracy – 87.71%**

**Overall Kappa Coefficient – 0.837**
over 87.71%, and the NLCD classification had an overall accuracy of 68.44%, 19.27% lower than the NAIP. The overall kappa coefficient for the NAIP classification was 0.837, and the overall kappa coefficient for the NLCD classification was 0.605. \textit{Landis and Koch, 1977} describe a kappa between 0.41 and 0.6 as moderate, 0.61 to 0.80 as substantial, and 0.81 – 1 as almost perfect agreement. Overall, the NAIP LULC classification was more accurate and in agreement with \textit{in-situ} validation points than the NLCD LULC classification.

The largest accuracy difference between the NAIP and the NLCD classification was in the agriculture class, with NAIP having a 92.50% consumer’s accuracy and NLCD having a 37.78% consumer’s accuracy. Many forested pixels were misclassified as agriculture in the NLCD classification, and this led to its low forest producer’s accuracy (43.75%) as well. For projects concerning the CyanoHAB risk mapping of small farm ponds, this low agriculture LULC could be problematic. The NAIP and NLCD classification also had a large accuracy disparity in the developed LULC class, with 84.78% and 66.67% accuracies respectively. Accurate classification of developed regions is necessary to capture potential sources of anthropogenic nutrient loading in respect to increased runoff, fertilizer leaching, and sewage leaks. NAIP’s one-meter resolution picked up finer developed features that the thirty-meter NLCD could not resolve.

Figures 8 and 9 shows boxplots by size category for the NAIP and NLCD LULC classes and CCD values. Each dot represents one waterbody. The difference between the NLCD and NAIP boxplot values are noteworthy, as small NLCD waterbodies had 39.57%, 33.42%, and 11.12% median agriculture, forest, and developed percentages while NAIP had 15.60%, 51.21%, and 20% respectively. This means that the NLCD classifies much more land as agriculture than NAIP, and that inversely NAIP classifies much more land as forest than the NLCD. On average
NAIP also classified twice as much land as developed than NLCD. Although their LULC proportions differ rather drastically, both Figure 8 and Figure 9 paint a similar story. They both highlight that smaller waterbodies have higher CCD values than medium and large waterbodies, and they also have more agriculture and less forested land in their watersheds. This pattern is intuitive as agriculture lands eutrophy surface water, and forested lands filter and absorb nutrients.

### 3.02 Cyanobacteria Cell Density Risk Map

Figure 10 and 11 shows the R-squared and p-value relationship between CCD and the 17 explanatory variables. For the NLCD and NAIP relationships, there were 13 and 12 classes with a p-value less than 0.05 (statistically significant) respectively. However, that was too many variables to create a novel CyanoHABs risk mapping equation. As described in section 2.06, the best-AIC MLR variables with p-values above 0.05 were removed. This left the NLCD and NAIP models with five variables each, a more manageable number of variables to apply this methodology to other sites/years.

Table 6 highlights the final inputs and outputs for the NCLD and NAIP-based MLRs. Both models used the same 5 variables; waterbody area, percent agriculture, percent forest, percent developed, and winter maximum temperature. Both models were statistically significant with p-values ~ 0, and they had adjusted R-squared values of 0.33 and 0.25 respectively. CCD values using the NLCD-based model can be calculated using Equation 5:

$$\text{CCD}_{\text{NLCD}} = 2,453.4 \times (\% \text{ Agriculture}) + 2,129.3 \times (\% \text{ Forest}) + 2,050.2 \times (\% \text{ Developed}) + 7,872.8 \times (\text{Winter Max Temperature}) - 362,473 \times (\text{Area}) - 248,378.2$$ (5)
Figure 8: Boxplot relationships between NLCD LULC classes and CCD values by waterbody size category. Smaller waterbodies had higher CCD and percent agriculture values, and lower percent forest values than medium and large waterbodies.
Figure 9: Boxplot relationships between NAIP LULC classes and CCD values by waterbody size category. Smaller waterbodies had higher CCD and percent agriculture values, and lower percent forest values than medium and large waterbodies.
Figure 10: R-squared (y-axis) and p-value (x-axis) of the 17 explanatory variables with CCD. The LULC classes are based on the NLCD, and the vertical red line indicates a 0.05 p-value. There are 13 classes with a p-value less than 0.05 (statistically significant).
Figure 11: R-squared (y-axis) and p-value (x-axis) of the 17 explanatory variables with CCD. The LULC classes are based on the NAIP, and the vertical red line indicates a 0.05 p-value. There are 12 classes with a p-value less than 0.05 (statistically significant).
Table 6: Final MLR outputs for (A) the NLCD LULC based dataset and (B) the NAIP LULC based dataset. The final sample size values (n) represent the original 771 waterbodies with the outliers, defined by over ten times the mean Cook’s distance, omitted.

A) NLCD

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-248,378.2</td>
<td>3 x 10^-8</td>
</tr>
<tr>
<td>Area</td>
<td>-326,473.0</td>
<td>2 x 10^-16</td>
</tr>
<tr>
<td>Percent Agriculture</td>
<td>2,453.4</td>
<td>6 x 10^-13</td>
</tr>
<tr>
<td>Percent Forest</td>
<td>2,129.3</td>
<td>1 x 10^-10</td>
</tr>
<tr>
<td>Percent Developed</td>
<td>2,050.2</td>
<td>2 x 10^-9</td>
</tr>
<tr>
<td>Winter Maximum Temperature</td>
<td>7,872.8</td>
<td>6 x 10^-6</td>
</tr>
</tbody>
</table>

n = 754
Adjusted R-Squared = 0.3341
Model P-Value = 2 x 10^-16

B) NAIP

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-180,984.4</td>
<td>6 x 10^-5</td>
</tr>
<tr>
<td>Area</td>
<td>-277,036.7</td>
<td>2 x 10^-16</td>
</tr>
<tr>
<td>Percent Agriculture</td>
<td>1,540.1</td>
<td>4 x 10^-6</td>
</tr>
<tr>
<td>Percent Forest</td>
<td>1,151.4</td>
<td>2 x 10^-4</td>
</tr>
<tr>
<td>Percent Developed</td>
<td>1,126.1</td>
<td>2 x 10^-3</td>
</tr>
<tr>
<td>Winter Maximum Temperature</td>
<td>9,692.5</td>
<td>7 x 10^-7</td>
</tr>
</tbody>
</table>

n = 755
Adjusted R-Squared = 0.2497
Model P-Value = 2 x 10^-16
CCD values using the NAIP-based model can be calculated using Equation 6:

\[
CCD_{\text{NAIP}} = 1,540.1 \times (\% \text{ Agriculture}) + 1,151.4 \times (\% \text{ Forest}) + 1,126.1 \times (\% \text{ Developed}) + 9,692.5 \times (\text{Winter Max Temperature}) - 227,036.7 \times (\text{Area}) - 180,984.4
\] (6)

While the variable coefficients for Equation 5 and 6 vary slightly, they have the same signs for all five variables. One would expect that as percent agriculture, percent developed, and winter maximum temperature increase, and area decease, that CCD would increase. Smaller waterbodies often had monoculture watershed LULC characteristics, exposing them to intensified effects for their LULC surroundings. Small waterbodies also often tend to be shallow, which favors strong biannual water column mixing and nutrient resuspensions that lead to CyanoHABs (Roy et al., 2013, Beaver et al., 2014). The literature has established a strong connection between agrarian dominated watersheds commonly found in the southeastern United States and CyanoHABs (Loftin et al., 2016, Beaver et al., 2014, Ploeg et al., 2010), where nutrients run off into local waterbodies, creating eutrophic conditions. Highly urbanized watersheds have also been documented as leaching phosphorus through wastewater and lawn fertilizers (Scheffer and Rinaldi, 1997, Soares et al., 2013) that contribute to CyanoHAB proliferation. These positive regression coefficients are in line with the literature and confirm the importance of agriculture and developed surfaces to CyanoHAB risk.

However, both models also state that as percent forest increases, so do CCD values. This goes against what the literature suggests, as forested lands provide environmental services that reduce surface water eutrophication (Beaver et al., 2014). It is possible that waterbodies in heavily forested areas lack the level of human intervention that waterbodies with more...
agriculture and developed landscapes may receive. That would mean that they are not managed or regularly checked for CyanoHABs, and are more likely to have higher CCD values because of their isolation. Contrarily, less forested waterbodies would be more likely to have active CyanoHAB monitoring and/or a nutrient management plan. The positive coefficients for percent forest could also be an artifact of Georgia’s historic LULC patterns. Periods of land disturbance via agriculture, and reclamation via reforestation in the piedmont over the last 70 years have left legacy sediments and nutrients that would not be expected in old growth forests (Coughlan *et al.*, 2017; Maloney *et al.*, 2008). These residual nutrients and historically poorly managed soils may make waterbodies more susceptible to CyanoHABs, at least more so than the current day LULC. A classification representing old growth and reforested areas could be more appropriate given the historic context of modern forests in the Georgia Piedmont. Forest is also an umbrella class that encompasses many different types of forest. A primary distinction could be made between naturally growing forests and forests used for silviculture. Management of forests in state and national parks are going to differ drastically from those in managed tree farms. Managed tree farms are much more likely to receive nitrogen and phosphorus treatments, in addition to herbicides, that could make them act more like an agriculture LULC than a forest LULC.

The importance of winter maximum temperature over summer maximum temperature was not anticipated either. It can be rationalized that all areas of the Georgia Piedmont are at risk for CyanoHABs due to their hot summer temperatures, but perhaps it is the winter temperatures that are more important to predict CCD and CyanoHAB risk. Warmer winter temperatures could allow cyanobacteria to live and be photosynthetically active year-round, and trigger CyanoHABs earlier in the year when warmer temperatures and nutrients are introduced during the spring, typically referred to as Spring Bloom in southeast U.S. Warmer winter conditions lack the
“natural predation” of a colder, harsher winter and may create environments at elevated risk for CyanoHABs. Global warming and climate change are expected to disproportionately impact winter temperatures (Piao et al., 2010) and cause a surge in winter-time cyanobacteria biomass (Cremona et al., 2017), potentially explaining the importance of winter maximum temperatures.

Although the NLCD model had a higher adjusted R-squared value, it is important to note that its classification was less accurate than that of the NAIP. The coarser, less accurate classification map generated better results than the NAIP model, and while it seems counterintuitive to use a less accurate classification, the NLCD represents a more novel way to map CyanoHAB risk. The NLCD is already available nationally in GEE is 1/900th the file size of a NAIP classification, making it easier to store and process as well. For this reason and having a slightly higher adjusted R-squared value, it is recommended to use the NLCD classification for doing CyanoHAB risk mapping. For general LULC mapping, especially for mapping finer features, it is recommended to use the NAIP classification.

3.02.1 Model Diagnostics and Risk Grouping

A fundamental assumption of linear regression is a normally distributed dependent variable. Figure 12 shows a histogram, Q-Q plot, and summary statistics for CCD values in the 771 waterbodies. Although unrealistic negative CCD values were calculated for some waterbodies (i.e. minimum value in Figure 12), they can be interpreted as a zero CCD for practical applications. The bell-curve shape of the histogram and near 1:1 Q-Q plot, along with a Shapiro-Wilks test p-value of 6 x 10^{-16}, confirm that CCD has a normal distribution. the NLCD and NAIP models were then checked for residual normality, multicollinearity, homoscedasticity, and outliers, and NCCDI breaks were set for risk group assignment.
Figure 12: (A) Density curve, histogram, and summary statistics of CCD 80\textsuperscript{th} percentile values in the 771 waterbodies, and a (B) near 1:1-line Q-Q plot indicates distribution normality.
Figure 13 shows model diagnostics for the NLCD model. The model residuals are normal and there are not any outliers with a Bonferroni p-value < 0.05. However, the three NLCD-based LULC classes, percent agriculture, percent forest, and percent developed were collinear. This means that those three variables are linearly related and do not have unique explaining power, which makes it difficult to tell which of the three variables is influencing CCD. A variance inflation factor greater than three typically describes a multicollinear explanatory variable. Additionally, a Breusch-Pagan test revealed heteroscedasticity in the model, indicating uneven residual magnitudes across the range of CCD values (i.e. larger residuals in high CCD values compared to low CCD values). This can result in varied confidence in a model depending on the magnitude of the value being predicted. Figure 14 shows a density curve and histogram of the NCCDI risk group breaks. The NCCDI Jenk’s natural breaks were 0-0.40 for low, 0.41-0.66 for medium, and 0.67-1 for high risk designations. Equation 5 was applied back to the 771 waterbodies and they were grouped by the Jenk’s natural breaks. Of the 771 waterbodies 26.85% were low risk, 37.09% were medium risk, and 36.06% were high risk.

Figure 15 shows model diagnostics for the NAIP model. The model residuals are normal and there are not any issues with explanatory variable multicollinearity. However, there were three observation point residuals with Bonferroni p-values less than 0.05. Even after removing points with a Cook’s distance greater than ten times the mean, outliers could be negatively influencing the MLR and skewing the results. Individual influential observations can drastically alter results and may need to be removed in the future. Additionally, a Breusch-Pagan test revealed heteroscedasticity in the NAIP model. Figure 16 shows a density curve and histogram of the NCDDI risk group breaks. The NCCDI Jenk’s natural breaks were 0-0.39 for low, 0.4-0.63 for medium, and 0.64-1 for high risk designations. Equation 6 was applied back to the 771
Shapiro-Wilk Residual Normality Test:
p-value = 5 x 10^{-12}, normal

Multicollinearity Variance Inflation Factor > 3:
   Area: False
   Percent Agriculture: True
   Percent Forest: True
   Percent Developed: True
   Winter Maximum Temp: False

Breusch-Pagan Test for Homoscedasticity:
p-value = 0.0147, heteroscedastic

Outlier Test:
   No residuals with Bonferroni p < 0.05

Figure 13: (A) NLCD MLR model residual distribution, (B) Q-Q plot, (C) Cook’s distance with a horizontal red line at 10 times the mean, and (D) diagnostics testing regression assumptions.
Figure 14: Density curve, histogram, and percent of waterbodies by risk group breakdown by NCCDI for the NLCD LULC model. The green, yellow, and red vertical lines represent the maximum NCCDI values for the low, medium, and high-risk designations respectively (i.e. low risk are 0 ~ 0.4).
Figure 15: NAIP MLR model residual distribution, (B) Q-Q plot, (C) Cook’s distance with a horizontal red line at 10 times the mean, and (D) diagnostics testing regression assumptions.
Figure 16: Density curve, histogram, and percent of waterbodies by risk group breakdown by NCCDI for the NAIP LULC model. The green, yellow, and red vertical lines represent the maximum NCCDI values for the low, medium, and high-risk designations respectively (i.e. low risk are 0 ~ 0.4).
waterbodies and they were groups by the Jenk’s natural breaks. Of the 771 waterbodies 24.38% were low risk, 37.35% were medium risk, and 38.26% were high risk.

Figure 17 shows an example risk map based on Equation 6, with risk represented by watershed color. Although this figure does not provide substantial insight to a waterbody’s risk for CyanoHABs since its CCD has already been calculated using Sentinel 2A imagery, future risk maps like this can be produced utilizing this methodology. Since all five variables used in the NLCD and NAIP MLR models are easily updateable, a new map could be made every year on April 1st. A user would have to create an updated LULC maps themselves to generate the most current LULC conditions to go along with the new January – March maximum temperature averages, but in theory that map could be made to predict CyanoHAB risk ever year. With a new creation date every April 1st, that also allows a risk map to be made before the late spring, summer, and fall seasons when CyanoHABs are especially prevalent.

3.03 Cyanobacteria-Positive Waterbody Validation

The 36 CyanoHAB-positive waterbodies histogram and density curve can be seen in Figure 18. The CyanoHAB-positive waterbodies mean and median CCD values were roughly double that of the 771 study waterbodies. The median CCD values for the 36 CyanoHAB-positive waterbodies was 113,031.77 cells/mL, compared to 51,787.58 cells/mL for the 771 waterbodies. Of the 36 waterbodies, 34 had non-zero CCD values and 20 had CCD values over 100,000 cells/mL. Although this is not a large sample size, it brings some confidence to the CCD equation by demonstrating its ability to detect high CCD values for known CyanoHAB-positive waterbodies.
Figure 17: Example risk map made from NAIP LULC model MLR outputs, applied back on the 771 watersheds. The main map resolves mostly larger low risk maps, and the subset primarily shows medium and high risk in smaller watersheds.
Figure 18: Density curve, histogram, and summary statistics of CCD 80\textsuperscript{th} percentile values in the 36 CyanoHAB-positive Georgia Piedmont waterbodies.
3.04 Cyanobacteria Cell Density World View Dashboard

Over the course of this project, GEE was an integral part of the data analysis and image processing workflows. As GEE proficiency increased and capabilities of the GEE platform were realized during the project, the idea was hatched to create a CCD world view dashboard. The general idea was to create a user-friendly interface that could map chl-a, CCD, and suspending sediment concentrations (SSC) anywhere in the world for the entire Sentinel 2A image collection. A user can either select a pre-determined waterbody from a dropdown menu, or zoom to their location of interest. The user then selects a cloud cover threshold, and a dropdown menu populates with all qualifying Sentinel 2A images of that area. The user can select a variety of image visualizations, including true color (bands 4/3/2), color infrared (bands 8/4/3), atmospheric (bands 12/8/4), NDVI, NDWI, chlorophyll-a, CCD, and SSC. Chlorophyll-a, CCD, and SSC visualization are for water pixels only. There is also a function that allows a user to get individual pixel values for chl-a, CCD, and SSC. This dashboard was not the focus of this study, but was concomitantly created during the study and could be a vital tool for water managers to use in the future alongside CyanoHAB risk maps. It is still under development, and Figure 19 shows an example of the beta-dashboard.

3.05 Applications of Results in Sustainable Food Systems

This study presents two useful tools with direct application to CyanoHAB management and monitoring that can help create more sustainable food systems. CyanoHAB risk maps can identify waterbodies that should be closely monitored, especially if they are commonly used as wading or drinking sources for livestock. Medium and high-risk waterbodies should be frequently monitored for bloom conditions, and immediate action should be taken if a
Figure 19: Screenshot of the cyanobacteria cell density world view dashboard. A user can select an area of interested, a cloud cover threshold, and click on water pixels to get chlorophyll-α, CCD, and suspended sediment concentration values.
CyanoHAB is occurring. Although the three-category risk designation system does not quantify risk, it can help focus or prioritize monitoring efforts, rather than having to check every waterbody. CyanoHAB risk maps can help streamline the monitoring process, allowing farmers to check for CyanoHABs without sinking inordinate amounts of time into visiting low-risk waterbodies.

Secondly, the cyanobacteria cell density world view dashboard puts remote sensing and spatial technologies into the hands of farmers and water managers themselves, without having to work through an image processing intermediary. This tool will work effectively in parallel with the risk maps, in that a farmer can check the CCD of a waterbody from anywhere that has an internet connection. This tool can act to focus monitoring efforts and not waste time checking non-affected waterbodies. Additionally, the cyanobacteria cell density world view dashboard is a completely free tool and can be easily customized by changing a few lines of code. The farmer themselves would not need to change the code; it can be manipulated by anyone and sent back to them remotely. Together, these tools can be utilized to identify at-risk waterbodies, limit livestock/CyanoHAB interactions, and contribute to a more sustainable food system.

3.06 Sources of Error and Uncertainty

The principal geographic units for each study site is its waterbody and watershed vector boundary. All subsequent calculations of explanatory variables and CCD are based on those boundaries, making their accuracy especially important. The discovery that only 771 of the original 2,100 delineated NHD defined “waterbodies” were confirmed surface water could warrant the creation of a separate database solely of surface waterbodies. While this study is
confident that the 771 waterbodies used were indeed permanent surface waters, future studies could benefit from creating their own waterbody vector boundaries. Additionally, smaller waterbodies that weren’t directly on a creek or stream did not produce very large dendritic watersheds. This could have been because they were closed waterbodies, meaning they aren’t connected to a linear water feature and only receive water recharge through direct rainfall and local topographic drainage. In that case, a traditional watershed delineation would be inappropriate and perhaps a simple buffering technique would be better. The separation or at least recognition of closed and open waterbodies could improve results.

While the NAIP LULC classification produced an “almost perfect agreement” according to Landis and Koch, 1977, the spectrally-limited four-band NAIP had speckled results when zoomed in close. Since NAIP is lacking the SWIR bands found in multispectral satellites like Sentinel 2A and the Landsat series, it had a difficult time classifying exposed soils and shadows. Shadows between tree canopies and cast by buildings were often misclassified as water, and bare agriculture fields were often classified as developed areas. These issues were exacerbated by NAIP’s one-meter resolution, in which its high-resolution was almost too high and classified the landscape’s LULC heterogeneity to a faulty degree. The NAIP imagery covering the Georgia Piedmont was also taken over the course of a summer 2015 (April – October 2015), rather than a single date. The resulting mosaic of these images had visible seamlines, which while not egregiously contrasting, affected pixel values and classifications on either side of the seamline. This temporarily large range made mapping agricultural lands difficult as well. Agriculture is a highly dynamic LULC class that changes from bare soil to fully vegetated multiple times per year; changes that occur very rapidly. Especially with NAIP’s four-band imaging, classifying these lands accurately was challenging.
Directly comparing classification accuracies and MLR results from NLCD and NAIP models was also problematic because of the four-year difference between the 2011 NLCD and 2015 NAIP classifications. The 2016 NLCD was not yet available at the time of this study, and the 2015 NAIP imagery was the most recently acquired for the Georgia Piedmont. It would have been ideal to compare products from the same time period, however, these were the available datasets. It can also be argued that the level I classification scheme used is too broad and failed to capture specific LULC classes contributing to a waterbody’s CyanoHAB risk. The LULC classes used are broad and general, and more specific classes could result in improved results for CyanoHAB risk mapping.

Errors may have also been introduced using the CCD model presented in Equation 4 (Page et al., 2016). This model was calibrated from a known CyanoHAB outbreak using 12 CCD data points provided by the Utah Department of Environmental Quality. This is a small sample size from a very large waterbody in a different climate than the Georgia Piedmont. To use remote sensing to definitively identify cyanobacteria, the satellite sensor needs to have a 620-nm band. This band is not available on the Sentinel 2A MSI, so CCD values carry the assumption that the algae bloom is a CyanoHAB, and not a green-algae bloom. Cyanobacteria has an absorption feature at 620-nm that green algae do not have.

The lower adjusted R-squared of the NAIP model versus the NLCD model was also not expected given the NAIP’s higher overall classification accuracy. Although it was more accurate, it could be possible CyanoHAB risk is more influenced by larger scale LULC patterns, especially given that one-meter LULC resolution could be excessive. Individual trees and developed structures may not as important of drivers as large forest patches and agrarian landscapes.
3.07 Future Work

The primary bottleneck during this study was the automated watershed delineation, as nearly all other computationally intensive analyses were performed in GEE. It took ArcGIS three days to delineate the original 2,100 waterbodies. At the time of this study, a method public to use GEE was watershed delineation was not available, and the ability to delineate watersheds in GEE would allow this study to have vastly expanded. Thousands of waterbodies could have been included if this workflow was realized, and will be necessary in the future to create regional, state, and/or national CyanoHAB risk maps. Additionally, waterbodies larger than one square kilometer should be included in future studies.

Given the higher accuracy of the NAIP classification but the higher adjusted R-squared value of the NLCD MLR CCD model, a “middle-ground” LULC classification should be investigated. The NAIP classification may be overly detailed and the NLCD classification is outdated and coarse, so perhaps a user-created ten-meter Sentinel 2A or pan-sharpened fifteen-meter Landsat 8 LULC classification would improve results. This would allow a user to create their own LULC classes and a most up-to-date classification image, rather than waiting two years for new NAIP imagery or five years for a new NLCD.

More work also needs to be done to validate and calibrate the CCD model with *in-situ* data. Finding publicly available *in-situ* CCD data can be difficult, but increasing the calibration and validation sample sizes will make the model more robust and universally applicable. Since the model assumes that all algal blooms are CyanoHABs, incorporating satellites with a 620 nm bands would help to confirm the presence of cyanobacteria in a waterbody. ESA satellites Sentinel 3 and Medium Resolution Imaging Spectrometer (MERIS) have 620 nm bands, however, they have a 300-m and 260-m spatial resolutions respectively. These satellites were not
used in this study because smaller waterbodies cannot be resolved at these resolutions. Even if only a few of these satellite’s pixels fall over a waterbody, they could be used to at least confirm a CyanoHAB, and then Sentinel 2A and Equation 4 could be used to confidently calculate CCD.

The 17 explanatory variables used in this study could be expanded and/or refined as new datasets become available, both on the web and in GEE. The 17 explanatory variables are in no way an exhaustive list of appropriate datasets that could help explain a waterbody’s CCD value and CyanoHAB risk. More variables and datasets should be analyzed as possible CyanoHAB risk predictors. It was also concluded that the use of aspect in this study was ineffective because it is not a ratio data type. Its values translate to degrees on a compass, and not a meaningful value that can be used in a MLR. Reshaping that data and incorporating similar topographic, hydrologic, and soils data would create more robust models. Additionally, the slope variable may have better represent the steepness of a watershed by using a measure of variance, not central tendency. If standard deviation was used in place of mean, better watershed steepness representation may have been attained. In the original proposal for this study, landscape configuration was supposed to be analyzed in addition to landscape composition. When the study migrated to GEE, studying landscape configuration became impractical and was ultimately abandoned. Landscape configuration needs to be considered in future studies, as it changes more frequently than landscape configuration and could better describe a waterbody’s immediate surroundings.

There also needs to be more work done on the cyanobacteria cell density world view dashboard. A fully-developed version of this tool would provide free, rapidly available water quality data to the public without having to download satellite images or learn complicated image processing techniques. Currently, only Sentinel 2A satellite imagery is being processed in
the dashboard, but if other satellites such as the Landsat series, MODIS, Sentinel 3, MERIS, and Planet were added, a wider variety of water quality parameters, and spatial and temporal resolutions would be available. In order to add these satellite series, atmospheric correction algorithms need to be developed, deployed, and tested for satellites without regularly produced surface reflectance products.

Finally, more research and development need to go into substantiating the use of “risk” in this study. Risk is a broad term with many definitions, and is especially discipline-dependent. Generally speaking, risk involves the uncertainty that an event will negatively impact a project or object (Project Management Institute, 1996), and/or the vulnerability of a population multiplied by their exposure to the hazard (White, 1974). The risk map and risk map equations created in this study are one dimensional and only consider predicted CCD values. To create a more holistic CyanoHAB risk map, many other factors needs to be accounted for, including demographic, socioeconomic, labor, and site specific environmental conditions. These datasets would help to identify vulnerable populations and hazardous working conditions that are equally important to creating a risk map as the predicted CCD values. Certain locales could have high CCD values but effective measures and procedures in place to handle CyanoHABs, while other with modest CCD values could be completely unprepared and expose people and animals to the many negative effects of CyanoHABs. This inclusion of these datasets would have future work even more interdisciplinary, and greatly enrich the overall meaningfulness of a CyanoHAB risk map.
CHAPTER 4
CONCLUSIONS

This study presented a streamlined watershed delineation technique, a one-meter LULC classification with an ~88% overall accuracy, and a simplified technique for using watershed-level variables to predict CyanoHAB risk in the Georgia Piedmont. While there are other batch watershed delineation processes currently available, this process required much fewer inputs and computational setup, and allowed watersheds to be delineated in a transparent fashion. Using ArcGIS ModelBuilder reduced the need for manual data manipulation to delineate watersheds, however, to perform studies with larger sample sizes, an integrated watershed delineation into GEE would be invaluable.

A one-meter 2015 NAIP LULC classification proved to be more accurate than the 2006 NLCD LULC classification according to all accuracy assessment metrics. With an overall accuracy of ~88%, the NAIP classification both quantitively and visually outperformed the ~61% overall accuracy of the NLCD classification. The NAIP provided much greater detail, especially with urban and water features, but was also limited by its four-band sensor. Intra-canopy space, tree shadow, and building shadow were often misclassified. Results should be reexamined after the release of the 2016 NLCD.

Using GRIDMET maximum winter temperature, percent agriculture, percent forest, percent developed, and waterbody area data, CyanoHAB risk were mapped using Equation 5 and Equation 6. Although the NAIP LULC classification had a higher accuracy, the NLCD MLR model had a 0.33 adjusted R-squared, compared to the 0.24 NAIP MLR adjusted R-squared.
Since the NLCD is available nationwide and necessitates less computational power, it is a better LULC class source for CyanoHAB mapping than NAIP. The five explanatory variables allow new CyanoHAB risks to be created every April, after the compiling of January-March average maximum temperature data. Updated LULC information could aid CyanoHAB risk mapping as LULC heterogenization occurs dynamically over short periods of time. More work needs to be done to calibrate and validate both the CCD and CyanoHAB risk equations to improve and bring about more robust analyses. Regardless, the CCD prediction and CyanoHAB risk mapping methodologies have the potential to be applied across the greater southeastern United States, and globally with the development of region-specific algorithms.

More research and development should be put into refining the cyanobacteria cell density world view dashboard, as it could be a vital tool for water resource managers alongside CyanoHAB risk maps. The two tools could work synergistically to identify at-risk waterbodies using the risk mapping techniques, and then monitoring at-risk waterbodies using the latest available satellite imagery. Having both a proactive geospatial tool in the risk map and reactive geospatial tool in the dashboard introduces new possibilities for CyanoHAB monitoring and mitigation efforts. The dashboard could also incorporate other satellites than Sentinel 2A, and calculate a spectrum of water quality parameters for a multi-sensor platform and approach to water quality mapping.
REFERENCES


Chen, Qiuwen, Han Rui, Weifeng Li, and Yanhui Zhang. 2014. Analysis of algal bloom risk with uncertainties in lakes by integrating self-organizing map and fuzzy information theory.


Elliot, J. Alex. 2010. The seasonal sensitivity of cyanobacteria and other phytoplankton to changes in flushing rate and water temperature. Global Change Biology 16: 864-76.


Havens, Karl, Hans Paerl, Edward Phlips, Mengyuan Zhu, John Beaver, and Akeapot Srifa. 2016. Extreme weather events and climate variability provide a lens to how shallow lakes may respond to climate change. Water (20734441) 8 : 1-18.


Haynie, Rebecca, Jamie Morgan, Brad Bartelme, Brad Willis, John Rodgers Jr., Lee Rogers, and Susan Wilde. 2013. Harmful algal blooms and toxin production in Georgia ponds. Georgia Water Resources Conference.


Catchment land use and trophic state impacts on phytoplankton composition: A case study from the rotorua lakes district, new zealand. *Hydrobiologia* 698 : 133-46.


Yellowstone Media Production. Yellowstone geysers - mammoth hot spring.