USING MACHINE LEARNING TO UNDERSTAND THE SPATIOTEMPORAL VARIABILITY OF HARMFUL ALGAE BLOOMS IN ILLINOIS WATERS

by

Supria Sarkar

M.S., University of Dhaka, 2018
B.Sc., University of Dhaka, 2016

A Thesis
Submitted in Partial Fulfillment of the Requirements for the Master of Science Degree

Department of Geography and Environmental Resources in the Graduate School
Southern Illinois University Carbondale
August 2021
THESIS APPROVAL

USING MACHINE LEARNING TO UNDERSTAND THE SPATIOTEMPORAL
VARIABILITY OF HARMFUL ALGAE BLOOMS IN ILLINOIS WATERS

by

Supria Sarkar

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of
Master of Science
in the field of Geography and Environmental Resources

Approved by:
Ruopu Li, Chair
Justin Schoof
Guangxing Wang

Graduate School
Southern Illinois University Carbondale
June 25, 2021
AN ABSTRACT OF THE THESIS OF

Supria Sarkar, for the Master of Science degree in Geography and Environmental Resources, presented on June 25, 2021, at Southern Illinois University Carbondale.

TITLE: USING MACHINE LEARNING TO UNDERSTAND THE SPATIOTEMPORAL VARIABILITY OF HARMFUL ALGAE BLOOMS IN ILLINOIS WATERS

MAJOR PROFESSOR: Dr. Ruopu Li

Harmful Algae Blooms (HABs) in inland waterbodies (e.g., lakes and ponds) pose serious threat to human health and natural ecosystem. Thus, it is imperative to assess HABs and their potential triggering factors over broader spatiotemporal scales. This study utilizes Chlorophyll-a (Chl-a) concentration in water samples collected from lakes in Illinois as an indirect measurement of HABs. The major objectives were to assess the spatiotemporal pattern of HABs over Illinois regions in recent decades, and to examine different machine learning models for predicting the Chl-a concentration based on publicly available water quality datasets. The Chl-a dataset was compiled from two different sources, the regular monitoring program by Illinois Environmental Protection Agency (IEPA) and the Volunteer Lake Monitoring Program (VLMP), the latter of which was primarily collected by citizen participants. Seven environmental and water quality zones were selected for spatial analyses. Additionally, the temporal patterns were assessed using time-series decomposition of monthly Chl-a concentration datasets. The machine learning pipeline includes two tasks: a regression modeling task for predicting Chl-a concentration, and a classification task for estimating lake trophic status. Different meteorological, land use and land cover, and lake morphometry variables were used as independent variables. Four regression models, i.e., Partial Least Squares Regression (PLSR), Support Vector Machine Regression (SVR), Artificial Neural Network
Regression (ANNR), and Random Forest Regression (RFR) were used for the first task of the modeling pipeline, and four classification models, i.e., Logistic Regression Classification (LRC), Support Vector Machine Classification (SVC), Artificial Neural Network Classification (ANNC), and Random Forest Classification (RFC), were used for the second task. Results indicate that: a) the Collinsville region in southwestern part of Illinois exhibited higher mean concentration of Chl-a in its lakes than any other regions from 1998 to 2018; b) the lakes with increasing trends in monthly mean Chl-a concentrations were also focused in the southwestern region; c) Random Forest outperformed all other models in both classification (Accuracy=60.06%) and regression ($R^2=38.88\%$); and d) the land use and land cover variables were found as the most important set of variables in Random Forest models.
ACKNOWLEDGMENTS

It is my genuine pleasure to express the deepest appreciation to my committee chair and advisor Dr. Ruopu Li for his outstanding and unparalleled support in conducting this research. His insightful guidance and suggestions assisted me to complete my master's thesis.

Additionally, I would like to thank my committee members, Dr. Justin Schoof and Dr. Guangxing Wang for their insightful comments and encouragement. Specifically, Dr. Schoof’s recommendations on meteorological variables extracted from different climate models were very helpful in my thesis. The spatiotemporal analysis section of this research was also enriched by the knowledge I acclaimed from Dr. Wang’s graduate courses.

I also want to acknowledge Tara Norris, Bureau of Water/Surface Water Section/Streams & Lakes Monitoring Unit, Illinois Environmental Protection Agency for providing us the water quality datasets.

Lastly, I would like to thank the entire faculty, staff and colleagues of the School of Earth Systems and Sustainability at the Southern Illinois University Carbondale for their help and guidance throughout my graduate study.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>i</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</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 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2. Research questions and objectives</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 2 LITERATURE REVIEW</td>
<td>5</td>
</tr>
<tr>
<td>2.1. HABs detection and monitoring practices</td>
<td>5</td>
</tr>
<tr>
<td>2.1.1. In-situ HABs assessments</td>
<td>5</td>
</tr>
<tr>
<td>2.1.2. Remote Sensing-based HABs assessments</td>
<td>7</td>
</tr>
<tr>
<td>2.2. Factors affecting HABs</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1. Manmade factors</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2. Natural factors</td>
<td>11</td>
</tr>
<tr>
<td>2.3. Relationship between HABs and lake trophic status</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER 3 MATERIALS AND METHODS</td>
<td>14</td>
</tr>
<tr>
<td>3.1. Study area and datasets</td>
<td>14</td>
</tr>
<tr>
<td>3.1.1. Study area</td>
<td>14</td>
</tr>
<tr>
<td>3.1.2. Datasets</td>
<td>15</td>
</tr>
<tr>
<td>3.2. Data preprocessing</td>
<td>17</td>
</tr>
<tr>
<td>3.2.1. Lake trophic status</td>
<td>17</td>
</tr>
</tbody>
</table>
3.2.2. Meteorological variables .......................................................... 19
3.2.3. LULC variables ........................................................................ 20
3.2.4. Lake morphometry variables .................................................. 20
3.3. Spatiotemporal assessment .............................................................. 23
   3.3.1. Spatial pattern analysis .......................................................... 23
   3.3.2. Temporal trend analysis ......................................................... 23
3.4. Machine learning pipeline ............................................................ 25
   3.4.1. Data preparation ..................................................................... 25
   3.4.2. Machine learning algorithms ............................................... 25
   3.4.3. Hyperparameter optimization .............................................. 33
   3.4.4. Model evaluation ................................................................. 34
CHAPTER 4 RESULTS ............................................................................ 37
4.1. Descriptive statistics of variables .................................................. 37
   4.1.1. Independent variables .......................................................... 37
   4.1.2. Dependent variables ............................................................ 39
4.2. Spatiotemporal assessment of HABs .............................................. 40
   4.2.1. Spatial pattern of Chl-a concentrations ................................ 40
   4.2.2. Chl-a temporal trends for lakes ........................................... 43
4.3. Machine learning models ............................................................... 47
   4.3.1. Selected features ................................................................. 47
   4.3.2. Chl-a prediction ................................................................. 48
   4.3.3. Lake Trophic Status classification ...................................... 50
   4.3.4. Variable importance in the random forest model ................. 53
CHAPTER 5 DISCUSSION ................................................................. 56
5.1. Spatiotemporal pattern of HABs outbreaks .................................. 56
5.2. Performance of machine learning models .................................... 57
5.3. Characteristics of IEPA and VLMP datasets ................................. 59
5.4. Limitations of the study .......................................................... 59
5.5. Policy recommendations .......................................................... 60
CHAPTER 6 CONCLUSIONS .............................................................. 63
REFERENCES .................................................................................. 65
APPENDIX A IEPA AND VLMP DATA QUALITY STANDARD .................... 77
APPENDIX B METEOROLOGICAL VARIABLE EXTRACTION .................. 79
APPENDIX C LULC DATA EXTRACTION FROM NLCD .......................... 81
APPENDIX D LAKE MORPHOMETRY VARIABLE EXTRACTION ............. 84
APPENDIX E DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES .. 85
VITA ......................................................................................... 86
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1: Datasets used in this study and their sources.</td>
<td>16</td>
</tr>
<tr>
<td>Table 3.2: Thresholds of TSI and Chl-a for determining trophic status.</td>
<td>18</td>
</tr>
<tr>
<td>Table 3.3: List of independent variables, their abbreviations used in this study, and associated units.</td>
<td>22</td>
</tr>
<tr>
<td>Table 3.4: Hyperparameters optimized in this study.</td>
<td>35</td>
</tr>
<tr>
<td>Table 4.1: Number of samples for each trophic status class in IEPA and VLMP.</td>
<td>40</td>
</tr>
<tr>
<td>Table 4.2: Number of lakes and their aerial characteristics for the seven regions.</td>
<td>41</td>
</tr>
<tr>
<td>Table 4.3: Information of lakes with different trend directions. All the listed lakes showed statistically significant trend at 99% confidence interval.</td>
<td>46</td>
</tr>
<tr>
<td>Table 4.4: Performance metrics of the test set for Chl-a prediction models.</td>
<td>48</td>
</tr>
<tr>
<td>Table 4.5: Performance metrics of the test set for lake trophic status classification models.</td>
<td>51</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1: Conceptualization of factors that trigger and accelerate algae blooms in waterbodies due to manmade interventions and climate changes. The figure was adapted from Paerl, Hall, et al. (2011) with appropriate permission.</td>
<td>10</td>
</tr>
<tr>
<td>Figure 2.2: Conceptual illustration of oligotrophic, mesotrophic, eutrophic, and hypereutrophic lakes. (VSWMS 2017)</td>
<td>12</td>
</tr>
<tr>
<td>Figure 3.1: Location of study area along with stations from IEPA and VLMP. Lake boundary and corresponding station locations are shown for three randomly selected lakes, i.e., Pinckneyville Reservoir (a), Du Quin Lake (b), and Sam Dale Lake (c).</td>
<td>15</td>
</tr>
<tr>
<td>Figure 3.2: Overall methods of the study. The approaches in (d) machine learning pipeline are fully described in Section 3.4.</td>
<td>17</td>
</tr>
<tr>
<td>Figure 3.3: Independent variable extraction process in this study. The gridMET cell is projected in a lake (a), the LULC from 2016 is overlayer with 3-km buffer of station locations (b), and the lake morphometry with the digital elevation model (c) for the sample lake.</td>
<td>19</td>
</tr>
<tr>
<td>Figure 3.4: ANN architecture for regression (a) and classification (b) used in this study. The major difference between the regression and classification architecture is the number of output neurons in the final layer. The number of neurons per layer is denoted on top of each layer.</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure 3.5: A schematic diagram of a RF example. There are three individual decision trees which together makes up a RF. The orange and blue circles are nodes of each tree which represents different classes. The direction sign in each tree shows how a RF decides from sample feature \((X)\) to target feature \((y)\). The splitting is done based on different metrics, e.g., mean squared error for regression or gini coefficient for classification. Figure adopted from: (Zhang et al. 2015) ................................................................. 31

Figure 3.6: Mechanism of k-fold cross validation where the training dataset is split into k- random subset and the model is trained k-times. The test set is kept completely aside for model evaluation. Figure credit: Pedregosa et al. (2011). ................................................................. 34

Figure 4.1: Descriptive statistics of independent variables from IEPA (a), VLMP (b). The X-axis represents the independent variables, and each component shows a corresponding violinplot. The blue value on the bottom of each violinplot represents the minimum value of that variable, whereas the red value on the top represents the maximum value. The variables which showed a significant difference between IEPA and VLMP in Student’s t-test (with 99% confidence interval), are marked as bold and green................................. 38

Figure 4.2: Frequency of trophic status classes for IEPA (a) and VLMP (b) along with the probability density distribution of Chl-a. The trophic
status classes were derived from the Chl-a values, which is represented by the bin size of different class. .................................................. 39

Figure 4.3: Spatial distribution of average Chl-a concentration values of individual stations (light green circles, the radius of circles represent the proportional intensity if Chl-a values), and average Chl-a of the regions in four timespans, i.e., 1998 – 2002 (a), 2003 – 2007 (b), 2008 – 2012 (c) and 2013 – 2018 (d). .......................................................... 42

Figure 4.4: Lake that had at least 50 entries of timeseries data and which showed positive (a) and negative (b) trend of Chl-a concentration .......... 44

Figure 4.5: The trend (blue line), associated direction of trend (grey dashed line), and aerial images of lakes that showed strong trend in both positive (a-c), and negative (d-f) directions. The positive trends were exhibited from Greenville Lake (a), Lake Jacksonville (b), and Crabapple Lake (c); whereas the negative trends were observed in Glen Jones Lake (d), Ginthers Lake (e), and Loch Lomond Lake (f). The associated p-value of each trend line are denoted. ............................. 45

Figure 4.6: Pearson correlation coefficient between each variable pairs for IEPA (a), and VLMP (b) datasets. The white color cells resulted p>0.01 in significance test and opted out from the analysis. The cells which showed a value greater than 0.90 are highlighted with a black square. ................................................................. 48
Figure 4.7: Scatterplots of measured Chl-a (µg/l) in X-axis and predicted Chl-a in Y-axis from IEPA (a-d), VLMP (e-h), and combined (COMB, i-l) datasets with PLSR, SVR, ANNR, and RFR algorithms. 49

Figure 4.8: Confusion matrix for classifiers trained with IEPA (a-d), VLMP (e-h), and combined (COMB, i-l) datasets. The classifiers were LRC, SVC, ANNC, and RFC. Each cell represents the percentage of true labels that were predicted for the corresponding cell. 52

Figure 4.9: Feature importance score extracted from the RFR (a-c) and RFC (d-f) algorithms for IEPA, VLMP, and combined (COMB) datasets. The LULC, meteorological, and lake morphometry variables are marked with green, red, and blue colors, respectively. 54
CHAPTER 1
INTRODUCTION

1.1. Background

Harmful Algae Blooms (HABs) are a significant and expanding threat to human health, fisheries, water safety, and tourism industries throughout the world (Brooks et al. 2016; Paerl et al. 2014). Typically, an ‘algae’ cell is a planktonic photosynthesizing organism, and over 5,000 species of aquatic phytoplankton have been described worldwide (Sournia et al. 1991). Among these species, cyanobacterial blooms are commonly denoted as harmful algae blooms or CyanoHABs, as it generates a toxic substance named ‘microcystin’ and plagues aquatic environments (Carmichael 2001). These blooms may produce hazardous toxic compounds, such as neurotoxins and hepatotoxins, which can cause visual disturbances, nausea, vomiting, gastroenteritis, liver failure, skin cancer, and even death (Greenfield et al. 2014).

HABs are becoming increasingly common in waterbodies and imposing economic and environmental threats at local, national, and global scales. For instance, in 2014, drinking water facilities had to be shut down due to an outbreak of Microcystis blooms in Lake Erie (Steffen et al. 2017). Frequent HABs at Lake Erie had resulted in vulnerable water supply systems for adjacent communities. Due to persistent occurrences of HABs, the states of Ohio and Michigan have declared the western part of the Lake Erie as impaired under the Clean Water Act (Davis et al. 2019). In the Lake Okeechobee of Florida, severe HABs have been occurring since the mid-1980s, and a state of emergency was declared for multiple counties in 2016 (Kramer et al. 2018). These blooms may have even shaped the results of state and federal elections in
Florida during the 2018 red tide events caused by *Karenia Brevis* (Stauffer et al. 2019). In the Gulf of Maine and Long Island Sound, continuous saxitoxin-producing *Alexandrium* blooms have been observed in recent years that had severe economic impacts (McGuire 2018). In terms of the global scale, massive fisheries closures were experienced in Chile due to the *Pseudoachattonella cf. verruculosa* and *Alexandrium catenella* blooms in 2016 (León-Muñoz et al. 2018). Similar blooms disrupted the fishing industry in coastal waters of Southeastern Australia from 2012 to 2017 (Trainer et al. 2020). Furthermore, HABs is associated with substantial economic loss. For instance, Dodds et al. (2009) reported that potential annual economic losses caused by eutrophication in U.S. freshwater systems amounts to $4.6 billion. In Florida alone, the HABs cause $22 million USD annual losses due to medical expenses and lost workdays, and $18 million due to disrupted commercial fishing industry (Kouakou and Poder 2019). Globally, HABs-related economic losses could approximate to around $10 billion USD annually (Bernard et al. 2014).

The influencing factors that trigger HABs in inland and coastal waters have been investigated in the literature. Studies have shown that excessive richness of nutrients (i.e., eutrophication) in water bodies (Heisler et al. 2008), increasing temperatures (Paerl and Huisman 2008), and over-harvesting of fish (Gobler et al. 2005) can be primarily responsible for severe HABs incidences. Specifically, the effects of anthropogenic eutrophication can be accelerated in inland water bodies with increased summer temperatures, frequent drought events, and heavy rainfalls (Ahn et al. 2002; Tyler et al. 2009). When a waterbody receives excessive amounts of Phosphorus (P) and Nitrogen (N) from upstream watersheds, the natural food-web is disrupted, resulting
in toxic cyanobacterial blooms (Paerl et al. 2014). Studies have found evidence that HABs may increase significantly with ongoing climate change. For example, increased CO\textsubscript{2} in coastal and inland ecosystems can exacerbate HABs and even increase the toxicity levels caused by some HABs (Fu et al. 2010). However, very few studies have considered a combination of climatic, land use, and other environmental factors in HABs outbreaks at regional scale.

Regional-scale spatiotemporal assessment of HABs requires accurate detection and monitoring of HABs across space and time, which is a technically challenging task. Chlorophyll-a (Chl-a) is the most common indicator for algae biomass estimation (Wiltshire et al. 1998). Hollister et al. (2016) utilized a Random Forest approach to estimate Chl-a and lake trophic status for the lakes in the United States using different analytes and geographic variables. Ho and Michalak (2020) examined associations between HABs and meteorological variables (such as, temperature and precipitation) to indicate total phytoplankton abundance, species dominance, and toxicity in the United States. Overall, such research efforts of using advanced machine learning algorithms to estimate HABs occurrence or intensity were limited to local scales (Yu et al. 2021; Li et al. 2014; Zhang et al. 2016) due to data availability. Limited data availability often prevents researchers and managers from properly understanding the factors that contribute to the regional variability of HABs occurrence over a large region.

1.2. Research questions and objectives

Similar to other places in the United States, HABs have been found a particular water quality concern in Illinois. Illinois Environmental Protection Agency (IEPA) had two regular HABs monitoring programs, the regular monitoring program by IEPA and the
Volunteer Lake Monitoring Program (VLMP), to collect various water quality parameters from 1998 until 2018. Such comprehensive HABs datasets allow water managers to understand the spatiotemporal patterns of HABs in different parts of Illinois over different periods of time.

The study addresses the following two research questions: a) **What are the spatiotemporal patterns of HABs in Illinois lakes?** and b) **What factors and how these factors contribute to the variability of HABs across Illinois?** To address the first question, the study utilized the Chl-a concentration as an indirect of HABs and performed spatial and time-series analyses. The second question was addressed by using three Chl-a datasets (i.e., the IEPA, VLMP and their combination) to develop machine learning models for predicting HABs.
CHAPTER 2
LITERATURE REVIEW

2.1. HABs detection and monitoring practices

In the past few decades, numerous HABs monitoring approaches have been carried out as either in-situ HABs detection, lab-based analysis or remote sensing-based assessment. Both methods have their own advantages and disadvantages in terms of accuracy, precision, cost of labor, and time. Research and management communities have been practicing these techniques to accurately detect HABs and reduce its adverse impacts for a long time. Following subsections discuss how the scientific community employs these techniques.

2.1.1. In-situ HABs assessments

Detecting HABs from in-situ water samples can be broadly classified into sensor-based, and laboratory-based assessments. The sensor-based measurements can be performed by a range of instruments, for instance, single to multispectral fluorometers, absorption/backscatter sensors, and particle size analyzers (Lombard et al. 2019). Fluorometer is capable of tracking fine-scale variations in HABs density, and community compositions. The device is effective in quick detection of HABs and have superiority in terms of size, power requirements, and cost (Zieger et al. 2018). Another example of sensor-based device is the Environmental Sample Processor (ESP), which uses immuno- or molecular probe-based assays to detect and quantify HABs species and toxin (Carvalho 2020). Recently, ESPs have been deployed on the United States West Coast (Yamahara et al. 2015), the Great Lakes area (Bowers et al. 2018), and the Gulf of Maine (Canfield 2013) to determine the feasibility of near-real time detection of
microcystin concentrations in waterbodies. A major advantage of these in-situ sensing techniques is that these devices can detect potential HABs in real-time with high accuracy and precision. In addition, such sensors can be incorporated into mobile, autonomous platforms, including autonomous underwater vehicles (AUVs) and autonomous surface vehicles (ASVs), which increases the spatiotemporal range of these instruments (Robbins et al. 2006; Schofield et al. 2008).

In addition to sensor-based in-situ assessments, there also exists laboratory-based analytical methods to detect HABs. The laboratory-based analytical approach can be divided into molecular and chemical approaches. Ohio Environmental Protection Agency (OEPA) adopted molecular approach for routine monitoring of HABs using CyanoDTec kit (Chiu et al. 2017). In 2016, the routine monitoring showed 100% correspondence of microcystin toxin detection above a 1.6 mg/L threshold with microcystin genes (OEPA 2018). On the other hand, chemical approaches have been available to some monitoring agencies for some time, but the results are semi-quantitative that are appropriate for only initial screening (USEPA 2017). However, some newly developed chemical analysis kits have been available recently that showed significant improvement in results. For example, the Mbio Toxin System (MBIO Diagnostics, Boulder, CO) has proved to be an effective kit to detect a range of freshwater and marine HABs toxins (McNamee et al. 2014). Another developing technology, Beacon Field Tube Kits (Beacon Analytical Systems Inc., Saco, ME) provides rapid quantitative in-situ assessment of samples without electricity need and specialized equipment (Stauffer et al. 2019).
Although high-throughput and efficient methods are available for measuring sample HABs toxicity from water samples, the major disadvantage of in-situ measurements is intensive labor and other costs. Therefore, it is not economical or efficient to get large-scale spatiotemporal HABs toxicity data using in-situ measurements. For large-scale data, the HABs toxicity has to be assessed using other proxy analyte information from water samples (Hollister et al. 2016).

2.1.2. Remote Sensing-based HABs assessments

Remote sensing approaches have been widely used to monitor toxic HABs over large areas in both freshwater and marine water systems (Anderson 1997). HABs affect the watercolor by increasing light backscattering with spectrally localized radiance from generic (Chlorophyll-a) and species-specific algae pigment absorption, for instance phycobiliproteins for cyanobacteria, fucoxanthin for diatoms and peridinin for dinoflagellates (Blondeau-Patissier et al. 2014). Therefore, Chlorophyll-a (Chl-a) can be used as a direct proxy for cyanobacterial biomass, and remote sensing techniques can be used to detect the blooms at the surface of water very efficiently (Cullen 1982). The remote sensing approaches can be broadly divided into space-borne and air-borne systems.

Satellite-based (space-borne) remote sensing techniques provide larger coverage of the study area and large-scale assessment scopes. The primary idea behind satellite based HABs detection technique has been the use of watercolor to quantify chlorophyll biomass and organic carbon in the upper water column from the satellite imagery (Klemas 2011). These imageries can either have multispectral (e.g., more than 3 bands) or hyperspectral characteristics (e.g., hundreds of bands). There
exist several satellite systems that provide different spatial, spectral, and temporal resolution images, such as, Sea-viewing Wide Field-of-view Sensor (SeaWiFS), Moderate Resolution Imaging Spectroradiometer (MODIS), Medium Resolution Imaging Spectrometer (MERIS), Visible Infrared Imaging Radiometer Suite (VIIRS), and the Ocean and Land Color Instrument (OLCI) sensor on Sentinel-3. The major advantage of using satellite-based remote sensing is the large spatial extent for analysis. However, limitations of satellite remote sensing include- 1) information is restricted to the surface of water (McManus et al. 2008), 2) strong interference with cloud cover and other atmospheric condition occurs, specifically near the coastal region that downgrades the cumulative information gathered from satellite images (Ryan et al. 2008), and 3) spatial resolution of satellite-based images are often not optimal for detecting ss (Wolny et al. 2020). The efforts to better resolve such issues have been well-studied from the last decade. Mouw et al. (2017) reviewed the algorithms that deconstruct watercolor data into HABs functional groups based on the advantages and limitations of different methods. The findings indicated that at least some of the methods can differentiate among larger and smaller size fractions of HABs communities.

Compared to space-borne sensors, lower altitude airborne and hyperspectral imaging (HSI) sensors can fly beneath clouds, whereas passive satellites cannot receive information under cloud. Airborne remote sensing also offers hundreds of narrow bands (10 – 20 nm) compared to the limited bands from satellite sensors (Bresciani et al. 2018). With benefits derived from lower altitude flights, HSI sensors provides an information-rich diversity of spectral signatures that can be used to sense a wide range of surfaces, habitats, taxonomic groups, and changes over time in these
signatures (Stauffer et al. 2019). Recent development in Unmanned Aerial Vehicles (UAVs) has also offered tremendous opportunity since the availability of low-cost UAVs and sensors are increasing. UAVs can now mount light-weight multispectral and hyperspectral sensors that can be very useful in detecting and monitoring HABs with higher spatial resolution (Pyo et al. 2018; Wu et al. 2019). Several studies have been carried out that utilized advanced machine learning and deep neural network architectures in HABs detection from UAV images. For example, Bollard-Breen et al. (2015) explored the capabilities of near-infrared (NIR) camera mounted in fixed-wing UAV and detected HABs with higher accuracy. Aguirre-Gómez et al. (2017) used simple red-green-blue (RGB) camera in a UAV to detect HABs in Chapultepec Park Lake, Mexico. Recently, Pyo et al. (2019) implemented Convolutional Neural Network (CNN) architecture and found significant accuracy in predicting phycocyanin and chlorophyll-a concentration of HABs in waterbodies.

Remote assessment of HABs monitoring using different sensors is very promising since it involves less labor-intensive tasks. However, there exists a takeoff between satellite and airborne sensors, i.e., satellite sensors can cover larger spatial scale at lower spatial resolution, whereas the airborne sensors can cover smaller area at higher spatial resolution. Additionally, very few researches has been conducted to monitor HABs at large spatiotemporal scale for inland waterbodies. Most of the literature was confined to study HABs in the ocean.

2.2. Factors affecting HABs

Many factors can significantly impact the timing and progress of HABs occurrences in freshwater and coastal waters. However, the current study focuses the
HABs occurrences in the inland freshwater systems. The factors can be broadly divided into two categories, i.e., manmade factors, and natural factors. Figure 2.1 illustrates a conceptual figure on how different manmade and natural factors play role in driving HABs outbreak within a waterbody.

Figure 2.1: Conceptualization of factors that trigger and accelerate algae blooms in waterbodies due to manmade interventions and climate changes. The figure was adapted from Paerl et al. (2011) with appropriate permission.

2.2.1. Manmade factors

The manmade factors can be termed as “eutrophication,” which is considered as one of the most significant ones triggering HABs worldwide (Wong et al. 2007; McGillicuddy 2010). Eutrophication is the enrichment of waterbodies by influx of industrial, urban, agricultural runoff and intensive animal farming fertilizers that run through the watershed into the waterbody (Glibert 2017). The most common nutrients
triggering HABs are Nitrogen (N) and Phosphorus (P), typically in the forms of NH₄, NO₃, and PO₄ (Brookfield et al. 2021). These nutrients have the capabilities of increasing both toxic and non-toxic phytoplankton, and human activities have altered the relative accessibility of certain nutrients that works in favor of HABs (Anderson 1997). Other than eutrophication, manmade factors also include aquaculture, shipping industry, and oil spills. Zohdi and Abbaspour (2019) documented that shrimp culture can increase in phytoplankton amount in a waterbody. One of the major HABs outbreaks in the Persian Gulf in 2008 was found to be triggered by the ballast waters in the shipping route between Qatar and Kuwait (Bakhtiar et al. 2020). Additionally, wastewater discharges pathogens and nutrients into water that results in increasing eutrophication, HABs growth rate, and decreasing water-soluble oxygen (OSPAR 2009).

2.2.2. Natural factors

In addition to human factors, HABs can be also triggered by natural factors. Trombetta et al. (2019) found that there is a positive correlation between water surface temperature and phytoplankton number and generally the effective temperature range for HABs growth is 25 – 30 °C (Chang et al. 2003). Additionally, enormous climatic fluctuations (e.g., El Nino, La Nina), which happens every 3 – 5 years, can change the wind direction and increase temperature that may cause HABs outbreaks in oceanic waters (Phlips et al. 2020). Another meteorological factor is heavy precipitation, which increases nutrients runoff to the waterbodies and accelerates HABs (Ho and Michalak 2020). Other than the meteorological factors, Read et al. (2015) demonstrated the relationship between lake morphometrical characteristics (e.g., surface area, volume,
depth) and HABs outbreaks. This is because eutrophication can be indirectly explained by the morphometrical parameters of the lake that can link the HABs outbreak.

2.3. Relationship between HABs and lake trophic status

Lake trophic status indicates the eutrophication level of a given waterbody which is an indirect proxy for HABs prevalence and biomass content (El-Serehy et al. 2018). Rusak et al. (2018) demonstrated that the HABs biomass variability increased with trophic status among lakes while, within-lake biomass variation increased with increasing variability in wind speed. Lake trophic status can be classified into four major classes, i.e., oligotrophic, mesotrophic, eutrophic, and hypereutrophic. Figure 2.2 illustrates the conceptual overview of four different trophic status levels for a lake.

![Figure 2.2: Conceptual illustration of oligotrophic, mesotrophic, eutrophic, and hypereutrophic lakes. (VSWMS 2017)](image)

Oligotrophic lakes generally have deep bottom, cold temperature, high oxygen content, sandy soil, and firmed structure (Cole et al. 1990). Because of such characteristics, the nutrient levels are usually lower in oligotrophic lakes, which does not
support HABs or large populations of aquatic plants. *Mesotrophic lakes* contain moderate amounts of nutrients and possess healthy ecosystem with diverse populations of aquatic plants, algae, and fish. Therefore, such lakes may contain occasional algae blooms with minimal proportions (Dokulil 1991). On the other hand, *Eutrophic lakes* are high in nutrients, low in oxygen conditions, have warm water, and soft soil sediments. These characteristics are suitable for algae bloom and may accelerate towards HABs outbreak with higher proportions (Soranno 1997). Finally, *Hypereutrophic lakes* are waterbodies that show even extreme characteristics than that of eutrophic lakes. Usually, hypereutrophic lakes are located near urban and agricultural area where the probability of direct discharge of nutrients is highly likely (Tang et al. 2016). Therefore, when a lake is surpassed with the nutrients from wastewater treatment facilities, sceptic systems, and heavy fertilizer usage, it increases the likelihood of heavy HABs outbreaks especially when coupled with extreme weather conditions (Paerl et al. 2011).
CHAPTER 3
MATERIALS AND METHODS

3.1. Study area and datasets

3.1.1. Study area

The study area is the State of Illinois (Figure 3.1), which is the 25th largest state in the United States in terms of landmass. The state’s border is formed by three major rivers (about 1466 km in length), i.e., Mississippi River on the west, Wabash River on the east, and Ohio River on the south. Illinois has 172,103 miles of streams that connect four major Hydrologic Unit Code 10 (HUC10) watersheds. The state also has 91,400 lakes and ponds, among which 3,256 lakes have a surface area of six acres or more and more than 87,000 water bodies can be considered as ponds (IEPA 2021a). These inland waterbodies not only provide valuable recreational and ecological resources, but also serve as potable water sources for many communities. Other than surface water resources, 75% of the state’s total land area are agricultural farmlands, which covers 27 million acres in total (USDA-NASS 2020). These agricultural farmlands produce mostly corn and soybeans which account for 54% and 27% of total farm revenue, respectively (IAGR 2021). In terms of climatic conditions, Illinois has four distinct seasons with different conditions. According to the State Climatologist Office for Illinois (Angel 2020), the temperature shows high variability along a north-south gradient with average annual temperature ranging from 8°C (north) to 15°C (south), and highs ranging from 14°C (north) to 20°C (south). Similarly, average precipitation exceeds 48 inches a year in the south, compared to less than 32 inches in the north.
Figure 3.1: Location of study area along with stations from IEPA and VLMP. Lake boundary and corresponding station locations are shown for three randomly selected lakes, i.e., Pinckneyville Reservoir (a), Du Quin Lake (b), and Sam Dale Lake (c).

3.1.2. Datasets

The datasets used in the study were collected from multiple sources (Table 3.1). The HABs intensity was represented using the Chlorophyll-a (Chl-a) concentration measured in µg/l. The Chl-a data was collected from two different sources from Illinois Environmental Protection Agency (IEPA), i.e., a) regular agency monitoring programs (denoted as IEPA dataset hereafter), and b) citizen-engaged Volunteer Lake Monitoring
Program (VLMP). The IEPA dataset was directly monitored by the Bureau of Water of IEPA, where the water samples were collected from different stations across Illinois and then analyzed for different analyte information. The VLMP dataset was also managed by IEPA, however, the main difference between the two datasets is that the VLMP water samples were collected by volunteer citizens and sent to IEPA for lab analyses (Ratliff 2017).

Table 3.1: Datasets used in this study and their sources.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Timespan</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEPA water quality</td>
<td>1998 – 2017</td>
<td>IEPA</td>
</tr>
<tr>
<td>VLMP water quality</td>
<td>2001 – 2018</td>
<td>IEPA</td>
</tr>
<tr>
<td>Lake shapefiles</td>
<td>/</td>
<td>NHD+ (USGS 2019) and manual digitizing</td>
</tr>
</tbody>
</table>

The meteorological variables were collected from gridMET, which is a daily high-spatial resolution (~4 km) surface meteorological data covering contiguous United States since 1979. Additionally, the Land use and land cover (LULC) dataset was collected from the Multi-Resolution Land Characteristics Consortium (MRLC) program for the years of 2001, 2006, 2011, and 2016 as 30-m rasters. Finally, the lake shapefiles were extracted from the National Hydrography Dataset Plus (NHD+) and manual digitization if a lake is not included in the NHD+ dataset.
3.2. Data preprocessing

The data processing step was the most important step for this study since the initial IEPA and VLMP datasets include different water quality parameters, qualifiers, and other metadata information. Figure 3.2 illustrates the overall methods of the study. Most of the data cleaning steps were conducted using Pandas (McKinney 2010), which is a Python package to handle tabular dataset.

![Diagram of data preprocessing methods]

Figure 3.2: Overall methods of the study. The approaches in (d) machine learning pipeline are fully described in Section 3.4.

3.2.1. Lake trophic status

The IEPA and VLMP data were first cleaned based on available qualifier information, i.e., certain entries were removed if the analyte value did not meet the
assigned data quality standard. More information about data quality can be found at APPENDIX A. During the data collection and analysis, IEPA marked certain entries if the sample did not meet standard data quality criteria. The samples that did not meet the standard data quality were removed. The Chl-a values considered for this analysis were already corrected for Pheophytin-a. Each sample of Chl-a measurement included location information (i.e., latitude and longitude), and collection date. The Chl-a values were processed as monthly averages because most of the lakes were sampled once a month or even less frequent. The lake trophic status was determined by calculating the Trophic Status Index of Chl-a (\( TSI_{Chl_a} \)) using the Equation (1) developed by Carlson (1977).

\[
TSI_{Chl_a} = 30.6 + 9.81 \times \ln (Chl_a)
\]  

where, \( Chl_a \) is the value of Chl-a concentration in \( \mu g/l \). Although Carlson (1977) also developed TSI equations for other analytes (i.e., total phosphorus and secchi depth) to determine the trophic status of a lake, only Chl-a was considered in this analysis due to the limited availability of other analytes. The trophic status (i.e., oligotrophic, mesotrophic, eutrophic, and hypereutrophic) for each sample was then determined based on the TSI values suggested by (Ratliff 2017) in Table 3.2.

<table>
<thead>
<tr>
<th>Trophic Status</th>
<th>( TSI_{Chl_a} )</th>
<th>( Chl_a ) (( \mu g/l ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oligotrophic (OLIG)</td>
<td>&lt;40</td>
<td>&lt;2.5</td>
</tr>
<tr>
<td>Mesotrophic (MESO)</td>
<td>40-50</td>
<td>2.5-7.5</td>
</tr>
<tr>
<td>Eutrophic (EUTR)</td>
<td>50-70</td>
<td>7.5-55</td>
</tr>
<tr>
<td>Hypereutrophic (HYPE)</td>
<td>&gt;70</td>
<td>&gt;55</td>
</tr>
</tbody>
</table>
3.2.2. Meteorological variables

The meteorological variables were extracted using the Google Earth Engine (GEE) platform and its Python API. An automated Python script (APPENDIX B) was developed using two Python packages, i.e., *EarthEngine* (Gorelick et al. 2017) and *geemap* (Wu 2020) packages to perform this task on the cloud. The gridMET raster cells that intersect with Chl-a sampling locations were selected. In terms of temporal dimension, all records for the months that match the samples were considered. Then the data points were averaged over time to calculate monthly values for all the gridMET variables. The gridMET dataset contains 16 meteorological variables, among which 9 variables were considered, including precipitation (mm), maximum relative humidity (%), minimum relative humidity (%), specific humidity (kg/kg), shortwave radiation (W/m2), wind direction (degrees clockwise from the North), minimum temperature (K), maximum temperature (K), and wind velocity (m/s). Figure 3.3a shows a sample station location relative to the corresponding gridMET cell dimensions.

Figure 3.3: Independent variable extraction process in this study. The gridMET cell is projected in a lake (a), the LULC from 2016 is overlayer with 3-km buffer of station
locations (b), and the lake morphometry with the digital elevation model (c) for the sample lake.

3.2.3. LULC variables

The LULC variables were extracted from the MRLC’s National Land Cover Datasets available from the years 2001, 2006, 2011, and 2016. For each Chl-a station, a 3-km circular buffer was created to extract five major LULC classes, i.e., Water, Forest, Crop, Wetland, and Impervious surface. The reason behind choosing a 3-km circular buffer is that it strikes a balance in a scale of being neither smaller than immediate land parcels nor greater than surrounding watersheds (Cheruvilil et al. 2013; Hollister et al. 2016). Within each circular buffer area, the total proportion of each LULC classes were calculated in an automated Python script (APPENDIX C). However, the LULC data were available from 4 different years starting from 2001-2016, whereas the Chl-a data were available from 1998-2017 (IEPA) and 2001-2018 (VLMP) with almost having one monthly average data for each year. Therefore, the match between Chl-a stations and the LULC was based on a nearest-neighbor approach. For example, a water sample for the date of July 2015 would be matched with the LULC data for the year 2016. Figure 3.3b shows a sample with the associated buffer areas and LULC map from 2016.

3.2.4. Lake morphometry variables

The lake morphometry variables represented the morphological characteristics of the lakes. We selected lake geometry (i.e., shoreline distance or perimeter in m, and area in m², volume in m³, maximum, average, and standard deviation of depth in m), and shape compactness indices (i.e., isoperimetric inequality, and digital compactness measure). First, the lake polygon shapefiles were extracted from the National
Hydrography Dataset Plus (NHD+) database. The data gaps in NHD+ were amended by on-screen digitization on NAIP 1-m resolution aerial imagery in ArcGIS. After processing all the lake shapefiles, the shoreline and area were calculated. The volume was calculated based on a 10-m resolution USGS DEM. A zonal statistics tool using ArcGIS Model Builder (APPENDIX D) was developed to extract the basic statistics of the DEM within each lake shapefile, and the volume was calculated by multiplying the area and the sum of the depths within a lake. Similarly, the maximum, mean, and standard deviation of depth were calculated using the zonal statistics tool.

Compactness indices represent the degree to which a shape is compact (Gillman 2002). For example, a circle is the most compact shape compared to a rectangle. We chose two most of the common compactness indices, i.e., IsoPerimetric Inequality (IPQ) developed by Osserman (1978), and Digital Compactness Measure (DCM) developed by Kim and Anderson (1984). IPQ is defined as in Equation (2):

\[
IPQ = \frac{4\pi A}{P^2}
\]

where, \( A \) is area in m\(^2\) and \( P \) is perimeter of the shape in m. The value of IPQ ranges from 0 to 1, where a high value is more compact than a shape of lower IPQ. On the other hand, DCM is calculated as the ratio of a shape’s area and the area of the smallest circle that circumscribes the shape. Figure 3.3c shows the smallest circle circumscribing a sample lake. We measured the smallest circumscribing circle area by using the minimum bounding geometry tool in ArcGIS. The range of DCM is also from 0 to 1, where the higher value represents more compact shape.

A list of the independent variables along with the abbreviations and units is shown in Table 3.3.
Table 3.3: List of independent variables, their abbreviations used in this study, and associated units.

<table>
<thead>
<tr>
<th>Variable Group</th>
<th>Variable Name</th>
<th>Abbreviation</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteorological</td>
<td>Precipitation</td>
<td>PRE</td>
<td>mm</td>
</tr>
<tr>
<td></td>
<td>Max relative humidity</td>
<td>RMX</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>Min relative humidity</td>
<td>RMN</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>Specific humidity</td>
<td>SPH</td>
<td>kg/kg</td>
</tr>
<tr>
<td></td>
<td>Shortwave radiation</td>
<td>SRD</td>
<td>W/m²</td>
</tr>
<tr>
<td></td>
<td>Wind direction</td>
<td>WDR</td>
<td>°N</td>
</tr>
<tr>
<td></td>
<td>Min temperature</td>
<td>TMN</td>
<td>K</td>
</tr>
<tr>
<td></td>
<td>Max temperature</td>
<td>TMX</td>
<td>K</td>
</tr>
<tr>
<td></td>
<td>Wind velocity</td>
<td>WVS</td>
<td>m/s</td>
</tr>
<tr>
<td>LULC</td>
<td>Water</td>
<td>WAT</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>Forest</td>
<td>FOR</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>Crop</td>
<td>CRP</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>Wetland</td>
<td>WET</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>Impervious Surface</td>
<td>IMP</td>
<td>%</td>
</tr>
<tr>
<td>Lake Morphometry</td>
<td>Area</td>
<td>ARE</td>
<td>m²</td>
</tr>
<tr>
<td></td>
<td>Shoreline</td>
<td>SHR</td>
<td>m</td>
</tr>
<tr>
<td></td>
<td>Max depth</td>
<td>MXD</td>
<td>m</td>
</tr>
<tr>
<td></td>
<td>Mean depth</td>
<td>MED</td>
<td>m</td>
</tr>
<tr>
<td></td>
<td>SD of depth</td>
<td>SDD</td>
<td>m</td>
</tr>
<tr>
<td></td>
<td>Volume</td>
<td>VOL</td>
<td>m³</td>
</tr>
<tr>
<td></td>
<td>IsoPerimetric Inequality</td>
<td>IPQ</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Digital Compactness Measure</td>
<td>DCM</td>
<td>/</td>
</tr>
</tbody>
</table>
3.3. Spatiotemporal assessment

3.3.1. Spatial pattern analysis

The spatial pattern analyses were performed for the Chl-a concentrations from the combined IEPA and VLMP dataset. The reason behind combining the datasets was that the distribution of Chl-a values over Illinois would be more uniform than taking individual dataset. For assessing the areas exhibiting higher concentrations of Chl-a over time, the study area was divided into several regions. The regions were demarcated based on seven IEPA zones for Land and Water Pollution Regional Offices (IEPA 2021b). The regions are divided based on the county boundaries, and named as Rockford, DesPlaines, Peoria, Champaign, Springfield, Collinsville, and Marion. Additionally, the timeline (1998 - 2018) was also divided into four different periods, i.e., (a) 1998 – 2002, (b) 2003 – 2007, (c) 2008 – 2012, and (d) 2013 – 2018 for analysis. All Chl-a values available within a region were averaged for each period and visualized in a map to highlight which pollution region showed higher mean concentration of Chl-a. Also, all available values for each period were averaged for each station on the spatial pattern maps.

3.3.2. Temporal trend analysis

The temporal trend analysis was performed for the combined dataset of IEPA and VLMP. However, the dataset had inconsistent sample size over space and time. Therefore, stations showing at least 50 monthly Chl-a values from 1998 – 2018 were considered for the trend analysis and preprocessed as time series.

The time series data may exhibit a variety of patterns which can be better explained by splitting the time series into several components (West 1997). Therefore, a
A time series is divided into trend-cycle component \((T_t)\), seasonal component \((S_t)\), and remainder component \((R_t)\). Here, the trend-cycle component represents a long-term increase or decrease in the data, seasonality dictates if there is any pattern affected by seasonal factors or some periods, and the remainder component is the remaining local variation between the time series and the combination of trend and seasonal component (Hyndman and Athanasopoulos 2021). Time series decomposition is the process of extracting these components from a given time series, and an additive decomposition model was considered in this study to extract the trend. The additive model is simply the sum of all three components in Equation (3):

\[
y_t = T_t + S_t + R_t
\]

where, \(y_t\) is the time series data. The timeseries decomposition was performed for each extracted lake with at least 50 timestamp entries using statsmodels package of Python. The missing values in the timeseries were imputed using a forward-filling method. However, to represent the magnitude and direction of the overall trend for each station, the strength of trend and slope were calculated. The strength of trend was calculated using the formula developed by Wang et al. (2006) as Equation (4):

\[
F_T = \max(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t+R_t)})
\]

where, \(F_T\) is the measure of the strength of trend, which ranges from 0 to 1, and \(\text{Var}\) is the variance of the components. If the value of \(F_T\) is closer to 0 then the strength of trend will be weak, whereas a value closer to 1 will represent a strong trend. However, the value of \(F_T\) does not indicate the direction of the trend. Therefore, a linear regression line was also fitted into each time series and the direction of slope was noted
to indicate which stations were exhibiting increasing (i.e., positive sign) or decreasing (i.e., negative sign) trend.

3.4. Machine learning pipeline

3.4.1. Data preparation

The machine learning pipeline started with splitting the entire dataset into two sets using a 70%-30% split ratio, i.e., training set (70% of the data) and testing set (30% of the data). The splitting was done in a stratified fashion which ensured that all four classes were available in the two sets. Since the independent variables had different units, all the variables were scaled from 0 to 1 for both train and test set. Then Pearson’s correlation coefficient was calculated between each of the variable pair to detect multicollinearity. If the absolute value of the correlation coefficient between a pair of variables was greater than 0.90 with 95% confidence interval ($p < 0.05$), then one of the variables was discarded.

3.4.2. Machine learning algorithms

The regression modeling (i.e., predicting Chl-a) and classification (i.e., classifying lake trophic status) tasks were performed by eight sets of machine learning models, in which four sets were utilized for the regression, and the other four for classification. The regression algorithms were Partial Least Squares Regression (PLSR), Support Vector Machine-based Regression (SVR), Artificial Neural Network Regression (ANNR), and Random Forest Regression (RFR). The classification algorithms selected for this study were Logistic Regression Classification (LRC), Support Vector Machine-based Classification (SVC), Artificial Neural Network Classification (ANNC), and Random
Forest Classification (RFC). These algorithms have been found robust models for water quality research (Ahmed et al. 2019; Haghiabi et al. 2018; Muhammad et al. 2015).

a) Logistic Regression Classification: Logistic regression (LRC) is one of the most common methods for predicting a discrete variable from a given set of independent variables. Although LRC has been primarily developed to predict the presence or absence of a characteristic (i.e., binary classification), it can be extended to solve multi-class classification problem as per the task in this study. A multinomial LRC algorithm was adopted in this study which used cross-entropy loss and predict multinomial probability distribution to natively support multi-class classification problems. As the hyperparameters of the model, $l_1$ and $l_2$ regularizations were used to optimize the logistic regression function. The $l_1$ and $l_2$ regularized logistic functions minimize the cost functions specified in Equation (5) and (6), respectively:

$$\min_{w,b} \frac{1}{2} w^T w + \sum_{i=1}^{n} \log(e^{-y_i(x_i^T w+b)} + 1)$$  \hspace{1cm} (5)

$$\min_{w,b} \|w\|_1 + \sum_{i=1}^{n} \log(e^{-y_i(x_i^T w+b)} + 1)$$  \hspace{1cm} (6)

where, $X_i$, and $y_i$ are the input and target features for trials $i=1,2,3,\ldots,n$. With each trial, the weight matrix, $w$ and bias term, $b$ updates until the model converges.

b) Partial least squares regression: The Partial Least Squares Regression (PLSR) was solely used for the regression task. PLSR is a technique which reduces the number of independent variables to a smaller set of uncorrelated variables and then performs least squares regression on these uncorrelated variables (Frank and Friedman
The general underlying model of a PLSR function is based on Equations (7) and (8):

\[ X = TP^T + E \] (7)
\[ Y = UQ^T + F \] (8)

where, \( X \) and \( Y \) are the independent and target variables, respectively. \( T \) and \( U \) are the projection matrices of \( X \) and \( Y \), respectively, and \( P \) and \( Q \) are the orthogonal loading matrices. The \( E \) and \( F \) are the error terms where it was assumed to be independently and identically distributed random variables. The decompositions of \( X \) and \( Y \) are done in a way so the covariance between \( T \) and \( U \) is maximized. Therefore, the number of components was the only hyperparameter considered in this study for PLSR training, which indicates the number of uncorrelated variables to be accounted for.

c) Support Vector Machine: Support Vector Machine (SVM) was utilized for both classification (i.e., SVC) and regression (i.e., SVR) problem in this study. SVM is an algorithmic implementation from the statistical learning theory developed by Cortes and Vapnik (1995). The SVM creates optimal decision boundaries (commonly known as hyperplane) between datasets by solving a constrained quadratic optimization problem, which can be defined as Equation (9) for a linearly separable problem of binary classification:

\[ f_{w,b} = \text{sign}(w \cdot x + b) \] (9)

where, \( f_{w,b} \) is the SVM function, \( w \) and \( b \) are the weights and biases, respectively. The \( \text{sign}(\cdot) \) function returns the sign of the output and if the sign is negative, then the class belongs to the negative class or vice-versa. For linearly non-separable case, the
optimal hyperplane is required to satisfy the following constrained minimization as

Formula (10) and (11):

\[
\min_{\mathbf{w}, \epsilon} \left( \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{L} \epsilon_i \right)
\]

\[
y_i (\mathbf{w}^T \phi(x_i) + b) \geq 1 - \epsilon_i
\]

where, \( C \) and \( k \) are used to weight the penalizing variables \( \epsilon_i \). \( \phi(\cdot) \) is a nonlinear function which maps the input space into a higher dimensional space, and these nonlinear functions are often known as kernel. For this study, two kernels were optimized, i.e., linear kernel (as described in Equation 9) and Radial Basis Function (RBF) kernel in Equation (12):

\[
\phi(\cdot) = k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}
\]

where, \( k(x_i, x_j) \) is the RBF function, \( \|x_i - x_j\|^2 \) is the squared Euclidean distance between the two feature vectors, i.e., \( x_i \) and \( x_j \), and \( \gamma \) is a hyperparameter to be tuned.

Therefore, for SVM hyperparameter optimization, kernel type, the value of \( C \) and \( \gamma \) were cross validated for optimization. For further information on the derivation of the SVM optimizations and kernels, please refer to Kim et al. (2003).

SVM is a popular machine learning algorithm for different problem domains. Although the theoretical backbone of the SVM algorithm was primarily developed for generating a hyperplane to classify data, the principles can be transferred for a regression problem as well. Based on the training data points, the SVM attempts to develop a curve, which can be used to establish relationship between vector space and the position of the curve for prediction.
d) Artificial Neural Network: Artificial Neural Network (ANN) or Multi-layer Perceptron (MLP) is a highly optimizable and advanced machine learning algorithm used in many disciplines. The ANN is consisted of at least three layers of nodes, i.e., an input layer, a hidden layer, and an output layer. The number of neurons for the input layer and the output layer are the number of features (i.e., independent variables), and number of classes, respectively. The number of hidden layers and neurons per layer are arbitrary numbers selected by user. Each neuron uses a nonlinear activation function and utilizes backpropagation algorithm to learn unique patterns in the training data. Like the LRC algorithm, a cross-entropy loss function was used to perform the multi-class problems.

An ANN architecture (Figure 3.4) has been developed for the classification problem based on multiple trials and errors. The final architecture includes 4 hidden layers with 32, 64, 128, and 256 neurons on each layer, respectively. The number of neurons selected for each layer was inspired by the most common DNN architectures (e.g., LeNet, AlexNet, and VGG16). Each neuron in the hidden layer was passed through a Rectified Linear Unit (ReLU), whereas the output layer was gone through a Softmax activation function for the classification task and a linear activation function for the regression task (Figure 3.4). The network was optimized with a stochastic gradient-descent algorithm named Adam (Kingma and Ba 2014) since it has shown efficient learning performance in many fields.
30

Figure 3.4: ANN architecture for regression (a) and classification (b) used in this study. The major difference between the regression and classification architecture is the number of output neurons in the final layer. The number of neurons per layer is denoted on top of each layer.

e) Random Forest: Random Forest (Breiman 2001) is an ensemble classification method that train several classifiers and combine the results through a voting process (Gislason et al. 2004). It creates multiple Classification and Regression Trees (CARTs), where each tree is trained on a bootstrapped sample of the original training data and the output class is determined by a majority vote of the trees (Breiman et al. 1984). For instance, Figure 3.5 shows an arbitrary random forest which explains the idea of multiple decision trees constructing a forest of trees.
Figure 3.5: A schematic diagram of a RF example. There are three individual decision trees which together makes up a RF. The orange and blue circles are nodes of each tree which represents different classes. The direction sign in each tree shows how a RF decides from sample feature \( (x) \) to target feature \( (y) \). The splitting is done based on different metrics, e.g., mean squared error for regression or gini coefficient for classification. Figure adopted from: (Zhang et al. 2015)

During the training process, the Random Forest (RF) searches across a randomly selected subset of variables to determine a split for each node based on some metric. The type of metric is different for regression and classification tasks. For regression problem, usually the mean squared error or mean absolute error are used as splitting criterion, whereas Gini Coefficient or Information Entropy are used for classification task. The regression criterions are expressed in Equations (13) and (14):
\[
\text{Mean Absolute Error} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{13}
\]

\[
\text{Mean Squared Error} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{14}
\]

where, \( N \) is the number of samples, \( y_i \) is the actual value and, \( \hat{y}_i \) is the predicted value by the model for each data point \( i \). On the other hand, the Gini Coefficient (or Gini Index) and Information Entropy are expressed as Equation (15) and (16):

\[
\text{Gini Index} = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{15}
\]

\[
\text{Information Entropy} = \sum_{i=1}^{C} -p_i \log_2 (p_i) \tag{16}
\]

where, \( p_i \) is the probability of the class that is being observed in a certain bag of samples and \( C \) is the number of classes. Other than the splitting criterions, another important hyperparameter for RF algorithm is knowing the number of trees to be considered. Therefore, the number of trees (or number of estimators) were determined based on hyperparameter optimization technique and cross validation. Additionally, the maximum depth of trees was also considered as a hyperparameter during the optimization since this can reduce the overfitting issue of the RF algorithm.

One of the major advantages of RF algorithm is to assess the feature importance (or variable importance) from the splitting criterion. Since the splitting of nodes in each tree occurs based on either mean squared error or mean absolute error for regression task and Gini Coefficient or information entropy for classification task, the average values of these criterion can represent the overall variable importance in the
model (Breiman 2001). This characteristic of RF algorithm allows to diagnose the variables after model training. In this study, the feature importance scores were extracted from the best performing RF models.

3.4.3. **Hyperparameter optimization**

The hyperparameters of different models were optimized using the grid search cross-validation technique (LaValle et al. 2004). The grid search technique considers a set of user-defined hyperparameters and then trains multiple models using the cross-validation folds, from which the best performing model is selected for final model evaluation. A 5-fold cross-validation was used where the training set was randomly divided into 5 sets of data and 4 sets were used for training the data and the rest set was used for validation set in each iteration. Figure 3.6 shows a schematic diagram of how a typical 5-fold cross validation works. The selection of hyperparameter values were solely based on literature search for different applications of the algorithms. The hyperparameters which were optimized in this study are listed in Table 3.4 along with the values. The explanation of each hyperparameter from different models are discussed in Section 3.4.1.
Figure 3.6: Mechanism of k-fold cross validation where the training dataset is split into k- random subset and the model is trained k-times. The test set is kept completely aside for model evaluation. Figure credit: Pedregosa et al. (2011).

3.4.4. Model evaluation

The regression models were evaluated based on the root mean square error (RMSE) along with the coefficients of determination ($R^2$). The equations for calculating RMSE and $R^2$ are as follows:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-1}}
\]  

(17)

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]  

(18)
where, $y_i$, $\hat{y}_i$, and $\bar{y}_i$ represent the measured, predicted, and average measured Chl-a values of the test dataset for the observations of $i = 1, 2, \ldots, n$, and $n$ is the number of testing samples.

Table 3.4: Hyperparameters optimized in this study.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hyperparameters</th>
<th>Values used for Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLSR</td>
<td>Number of Components</td>
<td>1 – 17 (with 1 increment)</td>
</tr>
<tr>
<td>LRC</td>
<td>Penalty</td>
<td>L1-Norm, L2-Norm</td>
</tr>
<tr>
<td>SVR/SVC</td>
<td>Kernel</td>
<td>Linear, Radial Basis Function</td>
</tr>
<tr>
<td></td>
<td>Gamma</td>
<td>Scale (1/#Features * Data Variance), Automatic (1/#Features)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.001, 0.01, 0.1, 1, 10, 100, 1000</td>
</tr>
<tr>
<td>ANNR/ANNC</td>
<td>Batch Size</td>
<td>32, 64, 128</td>
</tr>
<tr>
<td></td>
<td>Learning Rate</td>
<td>0.1, 0.01, 0.001, 0.0001</td>
</tr>
<tr>
<td>RFR/RFC</td>
<td>Number of Estimators</td>
<td>10 – 10000 (with 100 increment)</td>
</tr>
<tr>
<td></td>
<td>Maximum Depth of Tree</td>
<td>3 – 100 (with 1 increment)</td>
</tr>
<tr>
<td></td>
<td>Splitting Criterion</td>
<td>Mean squared error and mean absolute error for regression; Gini coefficient and entropy for classification</td>
</tr>
</tbody>
</table>

On the other hand, the classification models were evaluated using the overall accuracy and F1 score. The overall accuracy (%) simply indicates the ratio of correctly classified samples and total number of samples in the testing set. However, overall accuracy often fails to capture the class-specific model performance. Therefore, the confusion matrix of each model was calculated and the F1 score was calculated using Equation (19):
where, Precision is the number of positive predictions divided by the total number of positive class values predicted, and Recall is the number of positive predictions divided by the number of positive class values in the test set. Since the classification problem in this study is multi-class, each class was considered as positive class, whereas the others as negative. Then the result was averaged together for a multi-class precision, and recall values, which eventually led to a F1 score. The F1 score conveys the balance between the precision and recall for the classifier.
CHAPTER 4

RESULTS

4.1. Descriptive statistics of variables

4.1.1. Independent variables

The total number of samples for IEPA and VLMP datasets were 8,618 and 2,854, respectively. The difference in sample size may be explained by the fact that the IEPA dataset was available for 1998 to 2017 whereas the VLMP data was for 2001 to 2018.

The meteorological, LULC, and lake morphometry variables (22 variables) were extracted separately for both IEPA and VLMP stations. The statistical distributions of all independent variables are illustrated in Figure 4.1 using violin plots. Each variable is represented by one violin plot, which indicates the probability density of the data smoothed by a kernel density estimator (Scott 1979). In terms of IEPA samples, only the RMN (relative minimum humidity) and WDR (wind direction) showed normal distribution. Most of the other meteorological variables showed skewed to the bottom (i.e., smaller values) distribution except for PRE (precipitation), where smaller values were dominant. All the LULC and lake morphometry variables showed skewness to the top (i.e., higher values), which indicates most of the samples in these variables were low values. Similar characteristics were found for VLMP samples (Figure 5.1b), except the RMN showed a skewed distribution. APPENDIX E shows the descriptive statistics (i.e., mean, median, range, standard deviation, minimum, and maximum) of each variable from IEPA and VLMP stations. Both Figure 4.1 and APPENDIX E suggest that the independent variables used in the modeling process exhibited high variability with outliers.
Figure 4.1: Descriptive statistics of independent variables from IEPA (a), VLMP (b). The X-axis represents the independent variables, and each component shows a corresponding violinplot. The blue value on the bottom of each violinplot represents the minimum value of that variable, whereas the red value on the top represents the maximum value. The variables which showed a significant difference between IEPA and VLMP in Student’s t-test (with 99% confidence interval), are marked as bold and green.
4.1.2. Dependent variables

The dependent variables for the regression and classification tasks were Chlorophyll-a concentration (µg/l) and lake trophic status respectively shown in Table 3.2. Figure 4.2 shows the probability density of the continuous variable (Chl-a) along with the histogram of discrete variable (lake trophic status), where the bin size corresponds to the threshold values defined in Table 3.2. Additionally, the numbers of samples for each trophic status in IEPA and VLMP data are provided in Table 4.1.

Figure 4.2: Frequency of trophic status classes for IEPA (a) and VLMP (b) along with the probability density distribution of Chl-a. The trophic status classes were derived from the Chl-a values, which is represented by the bin size of different class.
Table 4.1: Number of samples for each trophic status class in IEPA and VLMP.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Oligotrophic (OLIG)</td>
<td>223</td>
<td>103</td>
</tr>
<tr>
<td>Mesotrophic (MESO)</td>
<td>902</td>
<td>510</td>
</tr>
<tr>
<td>Eutrophic (EUTR)</td>
<td>5054</td>
<td>1548</td>
</tr>
<tr>
<td>Hypereutrophic (HYPE)</td>
<td>2439</td>
<td>693</td>
</tr>
</tbody>
</table>

The probability distribution of the continuous variable (i.e., Chl-a) showed a skewed to the right distribution for both IEPA and VLMP datasets. This indicated that higher number of samples were present for the smaller values rather than the larger values. However, when the Chl-a data was classified using the thresholds provided in Table 3.2, the highest number of samples were found for the eutrophic class (EUTR) in both IEPA and VLMP datasets. The second highest samples were for hypereutrophic (HYPE) class, followed by the mesotrophic (MESO) and oligotrophic (OLIG) classes. Therefore, the regression analysis predicted continuous variable with a skewed distribution, whereas the classification task classified a variable with imbalanced classes.

4.2. Spatiotemporal assessment of HABs

4.2.1. Spatial pattern of Chl-a concentrations

The number of lakes along with the total and average surface area for each region are provided in Table 4.2. The highest number of lakes were found in DesPlaines (n=165), which is in the northeastern part of Illinois and close the Lake Michigan. However, these lakes did not cover larger surface area (total lake area of 92.07 km²) compared to the lakes in Marion region (total lake area of 157.56 km²). Other than the total lake area, the average lake area was found larger for the Champaign (3.38 km²)
and Collinsville regions (3.25 km²). Therefore, larger lakes are present in Champaign and Collinsville regions compared to the other regions where average lake area varies from 0.49 km² to 2.25 km².

Table 4.2: Number of lakes and their aerial characteristics for the seven regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of Lakes</th>
<th>Total Lake Area (km²)</th>
<th>Average Lake Area (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Champaign</td>
<td>31</td>
<td>104.78</td>
<td>3.38</td>
</tr>
<tr>
<td>Collinsville</td>
<td>39</td>
<td>126.84</td>
<td>3.25</td>
</tr>
<tr>
<td>DesPlaines</td>
<td>165</td>
<td>92.07</td>
<td>0.56</td>
</tr>
<tr>
<td>Marion</td>
<td>70</td>
<td>157.56</td>
<td>2.25</td>
</tr>
<tr>
<td>Peoria</td>
<td>30</td>
<td>14.74</td>
<td>0.49</td>
</tr>
<tr>
<td>Rockford</td>
<td>26</td>
<td>43.59</td>
<td>1.68</td>
</tr>
<tr>
<td>Springfield</td>
<td>41</td>
<td>58.48</td>
<td>1.43</td>
</tr>
</tbody>
</table>

Among seven IEPA regions, Collinsville region shows the highest amount of Chl-a concentration in its lakes for all the timespans, i.e., 1998 – 2018 (Figure 4.3). Other than the mean Chl-a for the whole region, individual stations also showed larger circles (i.e., larger circles represent higher concentration of Chl-a) over time in Collinsville. The highest number of lakes were found in DesPlaines region (Table 4.2), which were concentrated near the northeastern part that is close to the Lake Michigan. However, the mean Chl-a concentration in this region was comparatively lower for all the timespans than the other regions. The most southern part of Illinois, the Marion region, shows decreasing mean Chl-a values over time. At the most recent time (i.e., 2013 – 2018), three regions in the norther part (i.e., Peoria, Rockford, and DesPlaines) and the Marion region in south showed the lowest concentration of Chl-a in their lakes.
Figure 4.3: Spatial distribution of average Chl-a concentration values of individual stations (light green circles, the radius of circles represent the proportional intensity if Chl-a values), and average Chl-a of the regions in four timespans, i.e., 1998 – 2002 (a), 2003 – 2007 (b), 2008 – 2012 (c) and 2013 – 2018 (d).
4.2.2. *Chl-a temporal trends for lakes*

The number of lakes that had at least 50 records with timestamps is 95 from the combined dataset of IEPA and VLMP monitoring lakes. Figure 4.4 shows the location of those lakes in terms of the direction of their Chl-a trend (i.e., either positive or negative) and strength of trend ($F_t$). A $F_t$ close to 1.0 represents a strong trend, and a lower value indicates a weaker trend. The trend direction is denoted as positive or negative which indicates increasing or decreasing trend, respectively. The lakes which showed strong positive trends (Figure 4.4a) were distributed over the study area, whereas the Peoria region did not show any lake with strong negative trend (Figure 4.4b). There were several lakes found in the southern part of Illinois, which showed strong negative trends overall. Note that the negative trend does not indicate that the absolute concentration of Chl-a, because the average Chl-a concentration for these lakes can be smaller.
Figure 4.4: Lake that had at least 50 entries of timeseries data and which showed positive (a) and negative (b) trend of Chl-a concentration.

Figure 4.5 illustrates the trendlines associated with six lakes from both positive and negative trend groups. Table 4.3 shows 10 lakes and their associated information which showed most positive and negative trend from 2002 to 2018. The lakes with strong positive trends tend to be concentrated in the Springfield and Collinsville region of Illinois according to Table 4.3. Additionally, the lakes with strong negative trends were mostly found in the Marion and DePlaines region.
Figure 4.5: The trend (blue line), associated direction of trend (grey dashed line), and aerial images of lakes that showed strong trend in both positive (a-c), and negative (d-f) directions. The positive trends were exhibited from Greenville Lake (a), Lake Jacksonville (b), and Crabapple Lake (c); whereas the negative trends were observed in Glen Jones Lake (d), Ginthers Lake (e), and Loch Lomond Lake (f). The associated p-value of each trend line are denoted.
Table 4.3: Information of lakes with different trend directions. All the listed lakes showed statistically significant trend at 99% confidence interval.

<table>
<thead>
<tr>
<th>Trend Direction</th>
<th>Lake Name</th>
<th>City</th>
<th>Region</th>
<th>Strength of Trend ($F_t$)</th>
<th>Area (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Greenville Lake</td>
<td>Greenville</td>
<td>Collinsville</td>
<td>0.921</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Lake Jacksonville</td>
<td>Jacksonville</td>
<td>Springfield</td>
<td>0.899</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td>Crabapple Lake</td>
<td>Clayton</td>
<td>Springfield</td>
<td>0.882</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Channel Lake</td>
<td>Antioch</td>
<td>Springfield</td>
<td>0.869</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>Dawson Lake</td>
<td>Dawson</td>
<td>Springfield</td>
<td>0.862</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Oakton Lake</td>
<td>DesPlaines</td>
<td>DesPlaines</td>
<td>0.857</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Lake Hillsboro</td>
<td>Montgomery</td>
<td>Springfield</td>
<td>0.854</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Lake Glenn Shoals</td>
<td>Irving</td>
<td>Springfield</td>
<td>0.851</td>
<td>4.44</td>
</tr>
<tr>
<td></td>
<td>Otter Lake</td>
<td>Girard</td>
<td>Springfield</td>
<td>0.841</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>Lake Bloomington</td>
<td>Hudson</td>
<td>Champaign</td>
<td>0.836</td>
<td>2.40</td>
</tr>
<tr>
<td>Negative</td>
<td>Glen Jones Lake</td>
<td>Equality</td>
<td>Marion</td>
<td>0.924</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Ginthers Lake</td>
<td>Orel</td>
<td>Marion</td>
<td>0.904</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Loch Lomond Lake</td>
<td>Mundein</td>
<td>DesPlaines</td>
<td>0.884</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td>Fox Lake</td>
<td>Lake Villa</td>
<td>DesPlaines</td>
<td>0.876</td>
<td>10.08</td>
</tr>
<tr>
<td></td>
<td>Grass Lake</td>
<td>Lake Villa</td>
<td>DesPlaines</td>
<td>0.854</td>
<td>6.40</td>
</tr>
<tr>
<td></td>
<td>Pisticake Lake</td>
<td>Johnsburg</td>
<td>DesPlaines</td>
<td>0.851</td>
<td>7.05</td>
</tr>
<tr>
<td></td>
<td>Lake Marie</td>
<td>Lake Villa</td>
<td>DesPlaines</td>
<td>0.832</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>Lake Catherine</td>
<td>Lake Villa</td>
<td>DesPlaines</td>
<td>0.801</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Staunton Reservoir</td>
<td>Staunton</td>
<td>Springfield</td>
<td>0.786</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Forbes Lake</td>
<td>Omega</td>
<td>Collinsville</td>
<td>0.753</td>
<td>2.12</td>
</tr>
</tbody>
</table>
4.3. Machine learning models

4.3.1. Selected features

The independent variables (or features) were tested if there existed any multicollinearity among them. Multicollinearity indicates that a pair of variables have high correlation among each other and one of them can be removed to decrease model complexity (Senawi et al. 2017). Pearson’s correlation coefficients between each variable from both IEPA and VLMP samples were calculated at 99% confidence interval (Figure 4.6). A total of 7 pairs of variables showed high correlations with each other (\(|R| > 0.90\)), including 3 pairs from meteorological variables (TMN-SPH, TMX-SPH, TMX-TMN), and 4 pairs from lake morphometry variables (SHR-ARE, VOL-ARE, VOL-SHR, MXD-MED). Therefore, only TMN from the meteorological group, and VOL and MXD from the lake morphometry group were kept for further analysis among the 8 highly correlated variables. Similar correlations have been observed for both IEPA and VLMP samples, which led to 17 independent variables in further modeling process. Additionally, the correlation coefficient between each pair of variables was almost similar for both IEPA and VLMP variables (Figure 4.6).
4.3.2. Chl-a prediction

The Chl-a prediction was a regression problem where three versions of datasets (i.e., IEPA, VLMP, and both combined) were trained using PLSR, SVR, ANNR, and RFR models. The performance metrics calculated from the test sets are provided in Table 4.4. Additionally, the scatterplots of measured and predicted Chl-a values from the test sets of each dataset are illustrated in Figure 4.7.

Table 4.4: Performance metrics of the test set for Chl-a prediction models.

<table>
<thead>
<tr>
<th>Models</th>
<th>(R^2) (%)</th>
<th>(\text{RMSE (µg/l)})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IEPA</td>
<td>VLMP</td>
</tr>
<tr>
<td>PLSR</td>
<td>11.98</td>
<td>9.13</td>
</tr>
<tr>
<td>SVR</td>
<td>15.40</td>
<td>17.83</td>
</tr>
<tr>
<td>ANNR</td>
<td>19.62</td>
<td>9.01</td>
</tr>
<tr>
<td>RFR</td>
<td>38.88</td>
<td>32.94</td>
</tr>
</tbody>
</table>
Figure 4.7: Scatterplots of measured Chl-a (µg/l) in X-axis and predicted Chl-a in Y-axis from IEPA (a-d), VLMP (e-h), and combined (COMB, i-l) datasets with PLSR, SVR, ANNR, and RFR algorithms.

The regression results suggest that the best performance was achieved by the Random Forest Regression (RFR) algorithm with all three datasets. However, the best performance was achieved by the VLMP dataset with RFR ($R^2=0.39$), followed by the COMB ($R^2=0.34$) and IEPA ($R^2=0.33$) dataset. The ANNR performed slightly better than SVR for the IEPA ($R^2=0.20$) and COMB ($R^2=0.14$) datasets, whereas the SVR slightly outperformed the ANNR with VLMP dataset ($R^2=0.18$). The PLSR yielded the least performance scores among all four models and all three datasets.
Although the RFR outperformed the other three machine learning models, the model still suffers from 70-78% normalized root mean squares error for all three datasets. Figure 4.7 suggests that the models could not predict the higher values of Chl-a from the test dataset since the maximum range of measured Chl-a was found 200-300 µg/l, but the maximum predicted Chl-a values were around 150 µg/l for RFR (Figure 4.7d, Figure 4.7h and Figure 4.7i). This case was found even higher for the other models, specially the PLSR, where the highest predicted Chl-a was found around 80-90 µg/l. The reason behind the outperformance of RFR for VLMP than IEPA could be that the number of samples were smaller for VLMP than IEPA. This resulted in smaller chances of errors for the model. Additionally, the models also failed to predict the extreme high values for most of the cases. For example, the maximum value of the measured Chl-a was around 200-300 for IEPA and VLMP, whereas the maximum value predicted by the models were around 120-150 for most cases (Figure 4.7). The SVR with COMB dataset incorrectly predicted negative values (Figure 4.7j).

### 4.3.3. Lake Trophic Status classification

The classification metrics (i.e., accuracy and F1 score) are provided in Table 4.5, and the confusion matrices from each classifier and dataset are illustrated in Figure 4.8. The confusion matrices show the percentages of correctly classified samples for each class, where values closer to 100% indicates the best case.

The RFC outperformed all other models for classifying lake trophic status. The best result for RFC was obtained from the IEPA dataset (Accuracy=65.43%), followed by the combined dataset (Accuracy=64.29%) and VLMP dataset (Accuracy=64.29%). The ANNC, SVC, and LRC performed very similarly in the classification metrics (Table
The small differences in the accuracy and F1 scores may result from the randomness introduced in each model. This is because LRC, SVC, and ANNC initialize with random weights and then the weights are updated with iterations.

Table 4.5: Performance metrics of the test set for lake trophic status classification models.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IEPA</td>
<td>VLMP</td>
</tr>
<tr>
<td>LRC</td>
<td>58.93</td>
<td>55.54</td>
</tr>
<tr>
<td>SVC</td>
<td>59.67</td>
<td>58.46</td>
</tr>
<tr>
<td>ANNC</td>
<td>60.05</td>
<td>58.81</td>
</tr>
<tr>
<td>RFC</td>
<td>65.43</td>
<td>64.29</td>
</tr>
</tbody>
</table>
Figure 4.8: Confusion matrix for classifiers trained with IEPA (a-d), VLMP (e-h), and combined (COMB, i-l) datasets. The classifiers were LRC, SVC, ANNC, and RFC. Each cell represents the percentage of true labels that were predicted for the corresponding cell.

The confusion matrices (Figure 4.8) provide important insights into which classifier performed well with what type of classes. Since all three datasets (i.e., IEPA, VLMP, and COMB) contained imbalanced samples for each class (Figure 4.8), confusion matrices of each classifier show an in-depth analysis at the class level. For instance, both LRC and ANNC could not identify any oligotrophic (OLIG) and mesotrophic (MESO) classes except for the ANNC with VLMP dataset (Figure 4.8g, 39.2% samples were correctly identified as MESO). Since the number of samples were the highest for eutrophic (EUTR) class, most of the correctly classified samples were
found for the EUTR class. Although LRC classified most of the EUTR classes (over 95%), it failed to classify other classes. In contrast, the RFC could classify some of the lower eutrophic classes compared to LRC, SVC, and ANNC. However, the RFC still could not accurately classify over 50% of the hypereutrophic (HYPE) samples.

4.3.4. Variable importance in the random forest model

One of the key features of the random forest classification or regression algorithm is the variable importance scores calculated during the model training. Since random forest is a set of decision trees, the variables (or commonly known as features) selected to divide a tree are calculated using mean decrease impurity score. Therefore, how each feature decreases the impurity can be averaged for a random forest model and that can provide important information on which variables were important throughout the model training. Figure 4.9 shows the variable importance scores calculated for each random forest classifier trained with each dataset.
Figure 4.9: Feature importance score extracted from the RFR (a-c) and RFC (d-f) algorithms for IEPA, VLMP, and combined (COMB) datasets. The LULC, meteorological, and lake morphometry variables are marked with green, red, and blue colors, respectively.

The importance ranking of the variables found for the classification (Figure 4.9a-c) and regression models (Figure 4.9d-f) generally match in their patterns. However, the differences are noticeable among datasets. For instance, the LULC variables (i.e., green labels in Figure 4.9) were found highly important in IEPA dataset, whereas the meteorological variables (i.e., red labels in Figure 4.9) were important in VLMP dataset. But when the two datasets were combined (i.e., COMB dataset), the LULC variables were found as highly important compared to meteorological variables. The reason
behind this disagreement could be that the number of samples was lower for the VLMP dataset, and the samples were collected at different locations and time compared to the IEPA dataset. Since three versions of datasets resulted in similar important variables, it is safe to say that the most contributing variables are LULC variables. Among the LULC variables, the forest (FOR) and impervious surface (IMP) were found as the most important variables. However, the lake morphometry variables (i.e., blue labels in Figure 4.9) were found as the least important variables for most of the models, among which IPQ yielded comparatively higher importance than others.
CHAPTER 5
DISCUSSION

5.1. Spatiotemporal pattern of HABs outbreaks

The study utilized the Chl-a concentration measured in µg/l as a proxy for assessing the HABs outbreaks. The Collinsville region located in the southwestern part of Illinois showed comparatively higher Chl-a concentration over the study periods. According to Figure 4.3, the monthly mean Chl-a concentration values in Collinsville lakes were mostly in the range of Eutrophic status (Chl-a > 55 µg/l). The Collinsville region is consisted of Bond, Clinton, Fayette, Madison, Marion, Monroe, Randolph, St. Clair, Washington counties, where agricultural farmlands dominate the landscape (IAGR 2021). A possible reason behind this higher concentration Chl-a values in its lakes could be overland runoff and excess fertilizers ending up in the lakes. Since the Collinsville region is within two major watersheds of Illinois (McConkey et al. 2011), i.e., Kaskaskia River Watershed, and Mississippi River Valley Watershed, the flow of nutrients from agricultural farmlands may contribute to the surge of Chl-a concentration. Additionally, the higher availability of Chl-a data in this region may have an effect in resulting comparatively higher concentrations of Chl-a. Since the average temperature is comparatively higher in the southern part of Illinois (Angel 2020), the eutrophication rate in this region may be affected. However, more data is required to further evaluate the validity of the reasons behind such increasing Chl-a concentration.

The trend analysis of lakes was only performed for selected waterbodies which had at least 50 entries of monthly mean Chl-a data from 1998 to 2018, and only 95 of them satisfied that criteria among 395 waterbodies. Thus, the timeseries analysis in this
study could not cover all waterbodies of the regions. Among the lakes with strong positive trends, most of them are clustered in the Springfield and Collinsville regions (Table 4.3), which may be associated with intensive agricultural activities.

5.2. Performance of machine learning models

A combination of meteorological, LULC, and lake morphometry variables can explain the lake trophic status as a proxy for HABs outbreaks in inland waters of Illinois. However, among the popular machine learning models in environmental science, Random Forest tends to produce the best model fit for predicting both Chl-a concentration and lake trophic status. The superiority of Random Forest for water resources research has also been well documented in the literature (Addor et al. 2018; Bae and Park 2017; Berezowski and Chybicki 2018; Cabrera et al. 2018; Hollister et al. 2016). For example, Hollister et al. (2016) predicted Chl-a levels and trophic status of lakes for the entire U.S. using the Random Forest model. Two reasons may be behind the superiority of Random Forest. First, Random Forests can work well with both continuous and categorical independent variables (Horning 2010). In this study, most of the independent variables were continuous, but the lake morphometry variables acted as categorical. This is because the lake morphometry variables (such as area, volume, shoreline, IPQ, and DCM) varied spatially but not temporally. Second, Random Forest can increase the validation accuracy by reducing the chance of overfitting imposed by single Decision Trees (Susan et al. 2019). Decision Trees are well prone to overfitting issue, where the model performs well with the training set but performs very poorly with the validation set. This is because, a Decision Tree splits a node based on every possible independent variable, whereas each tree in a random forest can pick only from
a random subset of independent variables (Cutler et al. 2007). This nature of Random Forest results in a diversification of trees, and thus reduces the chance of overfitting.

Artificial neural networks (ANN) have been very popular due to their applicability to a variety of problems in different disciplines and outstanding performance. However, the ANN underperforms in both classification and regression tasks. The ANN architecture was developed by performing several trials and errors, because there is no single formula to specify the number of layers and neurons in each layer. In addition, the ANN model is largely a “black-box” (Tu 1996), which prevents an explicit diagnosis of statistical relationships between variables and is often subject to overfitting (Lawrence et al. 1997).

The hyperparameter tuning of the machine learning models were the most time-consuming part throughout the study. Since the Random Forest requires a certain number of Decision Trees to be developed, the choice of appropriate number of Decision Trees is very important (Cutler et al. 2007). However, a wide range of options were selected for choosing the appropriate number of Decision Trees (Table 3.4) in the Grid Search Cross Validation technique discussed in Section 3.4.3, which required the highest time to train the Random Forests. For instance, the RFR took approximately 4 days to train in a computer with Intel Core i7 processor and 16 gigabytes of primary memory. Alternatively, the other models took much less time to perform the hyperparameter optimization tasks. The fastest models in terms of training time were the PLSR and LRC for regression and classification tasks, respectively.
5.3. Characteristics of IEPA and VLMP datasets

The datasets used in this study were collected from two different water quality monitoring programs managed by IEPA. The major differences between the IEPA and VLMP datasets are who and where the samples are collected. IEPA water samples were collected by IEPA staffs, but VLMP dataset was collected by volunteer citizens. After the sample collection, all the samples were sent to IEPA laboratories for Chl-a measurement. Therefore, the difference can be seen in the timestamps and locations of sample collection. The number of lakes found within IEPA and VLMP datasets were 349 and 183, respectively. Since IEPA maintained continuous monitoring of lake water quality from 2001 to 2018 as part of the regular maintenance of natural lakes and reservoirs, the number of samples and lakes were higher than the VLMP lakes. Additionally, the mean areas of IEPA and VLMP lakes were 1.70 and 0.90 km², respectively, which indicates that the VLMP focuses on smaller water bodies. It is likely that citizen scientists collected more water samples from smaller and/or private lakes and ponds that are only accessible to community members. The VLMP can effectively supplement the data gaps in regular IEPA’s HABs monitoring program.

5.4. Limitations of the study

Although the study shows the spatiotemporal pattern of HABs outbreaks in Illinois inland waterbodies and models the relationships between HABs and environmental variables, there were some limitations observed during the analyses. Major limitations include:

a) The Chl-a data from IEPA and VLMP sources were not uniformly distributed over space and time. Therefore, the temporal trend analysis in Section 4.2.2
relied on forward-filling method when certain dataset was not available in a time series. If more data were available, the timeseries decomposition would be more accurate and precise.

b) The HAB outbreak was assessed using a proxy indicator from water quality variables, i.e., Chl-a. However, if direct measurements of Microcystin (the toxic element present in cyanobacterial blooms) were widely available, the assessment would better target HABs outbreaks. Although the IEPA dataset contains Microcystin measurements, the number of such samples is very low compared to the Chl-a samples.

c) The Chl-a concentration from each sampling station was considered as representative of water quality for the entire lake. Whenever multiple stations were available for larger lakes, the samples were averaged if being collected in the same dates. Thus, the Chl-a samples used in the analyses cannot represent the spatial variability in the same lake without bias.

d) Machine learning models were trained using a dataset that had skewed distribution of dependent and independent variables (Figure 4.1).

5.5. Policy recommendations

The results suggest that HABs in inland waterbodies are a common phenomenon in Illinois and continuous water quality monitoring can help better understand the relationships between natural processes and algae growth in lakes. Therefore, continuous lake monitoring through either citizen participation or data collection by experts is imperative to ensure the health of Illinois waters. The following recommendations are made to assist the policy makers:
a) The utilization of in-situ sensor-based HABs instruments (such as, Environmental Sample Processor, ESP, as discussed in Section 2.1.1) during the regular monitoring of lake water quality can provide direct measurement of HABs toxicity. Availability of such direct measurement of HABs can allow researchers to develop more accurate models that can predict the probability of HABs outbreak at certain waterbody for larger regions.

b) The VLMP has proved to be an effective program to assess lake water quality. However, the Illinois VLMP program has been suspended in 2019 until further notice (IEPA 2021c). The ILEPA may reinstate this program so that citizen volunteers can continuously participate in this important water quality monitoring program and provide water quality data from smaller water bodies.

c) Secchi disk depth is an inexpensive and simple-to-use instrument that can be indirectly connected to HABs intensity (Setiawan et al. 2019). The first tier of VLMP included this easy measurement into its protocol but due to the current suspension of the program, this dataset is not being collected anymore. Even if the entire VLMP program cannot be reinstated, at least the collection of Secchi disk depth should continue.

d) Lake water quality assessment agencies should also consider the use of Unmanned Aerial Vehicles (UAVs) with remote sensing sensors to model the spatial pattern of HABs within smaller water bodies. Recent availability of low-cost sensors and UAVs along with advanced machine learning and deep
learning techniques (Wu et al. 2019) will develop smart lake monitoring systems with high accuracy and precision.
CHAPTER 6

CONCLUSIONS

Using Chlorophyll-a as an indirect measure for the inland HABs in Illinois, this study assessed the spatiotemporal patterns of HABs and examined their relationships with a combination of environmental variables using several machine learning models. For spatiotemporal patterns, the Chl-a dataset from two different IEPA sources was analyzed using different geoprocessing tools and maps. For statistical relationships, different meteorological, LULC, and lake morphometry variables were explored. Commonly used machine learning algorithms, such as logistic regression, partial least squares regression, support vector machine, artificial neural network, and random forest, were used to predict the Chl-a concentration and classify the lake eutrophic status. The major findings are:

a) The Collinsville region exhibited higher concentrations of Chl-a in its lakes compared to other regions from 1998 to 2018 consistently. Comparatively, the northern regions, i.e., Peoria, Rockford, and DesPlaines have showed lower concentrations of Chl-a in the lakes.

b) The lakes with increasing trends were located mostly in the Springfield and Collinsville regions. Therefore, more policy-level focus should be given to these regions in terms of future HABs outbreaks.

c) Among the machine learning models, random forest outperformed all other algorithms in both classification (Accuracy=60.06%) and regression ($R^2=38.88\%$) tasks.
d) The IEPA data yielded the best performance metrics with Random Forests for both classification and regression problems. The VLMP dataset exhibited slightly lower accuracy metrics, which may be due to the lower sample size in VLMP dataset.

e) The LULC variables were found as the most important set of variables in Random Forest models. Forest, impervious surface, and croplands were found to be the most important LULC factors for explaining the variation of Chl-a concentration.

This study is probably among the first studies that examine HABs using comprehensive statewide HABs monitoring datasets. Future endeavors in studying HABs may include a comparative study in other states in United States. This effort may allow more insights into the machine learning models for predicting HABs. If Random Forest still yield superior performance in predicting HABs, the model may be standardized for a HABs early alert program. Such an early alert system may help decision makers and residents to be better prepared for safety use of recreational water bodies.
REFERENCES


cyanobacterial harmful algal blooms." *Harmful Algae* 81: 59-64. [https://doi.org/10.1016/j.hal.2018.11.016](https://doi.org/10.1016/j.hal.2018.11.016).


---. 2021b. "Land and Water Pollution Regional Offices." Springfield, IL, USA. [https://www2.illinois.gov/epa/about-us/Locations/Pages/regional-offices.aspx](https://www2.illinois.gov/epa/about-us/Locations/Pages/regional-offices.aspx).


Lawrence, S., C. L. Giles, and A. Tsoi. 1997. "Lessons in Neural Network Training: Overfitting May be Harder than Expected." AAAI/IAAI.


APPENDIX A
IEPA AND VLMP DATA QUALITY STANDARD

The sample value of Chl-a which met the qualifiers marked as bold were only selected for analysis in this study.

<table>
<thead>
<tr>
<th>Qualifier Code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><strong>Averaged Result.</strong> For a bacteria result, the reported value is the arithmetic mean of two or more colony counts in the acceptable range and at the same filtration volume or dilution.</td>
</tr>
<tr>
<td>B</td>
<td><strong>Used for bacteria results only.</strong> The reported value is based on colony counts outside the acceptable range.</td>
</tr>
<tr>
<td>B1</td>
<td><strong>The sample matrix caused possible effects on measurement. The result may be biased low.</strong></td>
</tr>
<tr>
<td>B2</td>
<td><strong>The sample matrix caused possible effects on measurement. The result may be biased high.</strong></td>
</tr>
<tr>
<td>C</td>
<td><strong>&quot;Value calculated. Reported result is calculated from other data.&quot;</strong></td>
</tr>
<tr>
<td>E</td>
<td><strong>For a bacteria result, the reported value is the arithmetic mean of two or more colony counts within the acceptable range but from different filtration volumes or dilutions.&quot;</strong></td>
</tr>
<tr>
<td>J</td>
<td><strong>Estimated value. Used only for results obtained in the field.</strong></td>
</tr>
<tr>
<td>J1</td>
<td><strong>Estimated value: Analyte was detected between the specified reporting limit (RL) and the method detection limit (MDL).</strong> For older records (i.e., approx. year 2003 and earlier) for which the &quot;Remark Code&quot; = &quot;J&quot;, the meaning of &quot;J&quot; is broader. For these records &quot;J&quot; is equivalent to qualifier codes &quot;J1&quot;, &quot;J3&quot;, &quot;J4&quot;, or any combination thereof.</td>
</tr>
<tr>
<td>J2</td>
<td><strong>Surrogate compound recovery limits have not been met</strong></td>
</tr>
<tr>
<td>J3</td>
<td><strong>Refer to case narrative for specific requirements.</strong></td>
</tr>
<tr>
<td>J4</td>
<td><strong>The reported value failed to meet the established quality control criteria for either precision or accuracy possibly due to matrix effects.</strong></td>
</tr>
<tr>
<td>J6</td>
<td><strong>The sample matrix interfered with the ability to make any accurate determination (i.e. oily samples, high mineral content in SOCs, etc.)</strong></td>
</tr>
<tr>
<td>J7</td>
<td><strong>Blank Spike failed high – possible high bias or false positive result.</strong></td>
</tr>
<tr>
<td>J9</td>
<td><strong>Blank Spike failed low – possible low bias or false non-detect result.</strong></td>
</tr>
<tr>
<td>K</td>
<td><strong>Used for field-recorded results only. Accuracy of the reported value is questionable.</strong></td>
</tr>
<tr>
<td>L</td>
<td><strong>&quot;Actual value not known but known to be less than value shown.</strong> For older records (i.e., approx. year 2003 and earlier) for which &quot;Remark Code&quot; = &quot;K&quot; or &quot;L&quot;, the value reported in the result field typically represents the minimum or maximum, respectively, reporting limit of the laboratory method used to obtain that result. Therefore, data users are cautioned against simply accepting and using these reporting-limit values as the actual result.**</td>
</tr>
<tr>
<td>L1</td>
<td><strong>For a bacteria result, this qualifier is used when no colonies are observed in any volume or dilution. The reported result is calculated as if the one colony were observed in the largest volume or dilution (the greatest amount of sample filtered).&quot;</strong></td>
</tr>
</tbody>
</table>
| L2             | **"The concentration of the analyte was above the acceptable level for quantitation.** For older records (i.e., approx. year 2003 and earlier) for which "Remark Code" = "K" or "L", the value reported in the result field typically represents the minimum or maximum, respectively, reporting limit of the laboratory method used to obtain that result. Therefore, data users are cautioned against simply accepting and using these reporting-
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>For a bacteria result, this qualifier is used when colony counts in all filtered volumes or dilutions are greater than the 200-colony maximum practical quantitation limit. The reported result is calculated as if 200 colonies were observed in the smallest volume (largest dilution), i.e., the least amount of sample filtered.</td>
</tr>
<tr>
<td>N</td>
<td>For Secchi-depth readings only. The recorded Secchi depth represents a water-column depth to the bottom, i.e., the Secchi disc was visible to the bottom of the waterbody.</td>
</tr>
<tr>
<td>Q</td>
<td>For Secchi-depth readings only. Actual Secchi depth is unknown, but greater than the recorded depth. The Secchi disc was hidden by aquatic plants when descending or ascending in the water column.</td>
</tr>
<tr>
<td>S</td>
<td>Presence of material verified (i.e., positive detection). Value is estimated.</td>
</tr>
<tr>
<td>TNTPC</td>
<td>Presumptive evidence of presence of material. The qualifier shall be used if: (1) The component has been tentatively identified based on mass spectral library search; or (2) There is an indication that the analyte is present, but quality control requirements for confirmation were not met.</td>
</tr>
<tr>
<td>V</td>
<td>Holding time exceeded.</td>
</tr>
<tr>
<td>X</td>
<td>Test results provided by outside source.</td>
</tr>
<tr>
<td>W</td>
<td>Used for bacteria results only. Too Numerous to Count</td>
</tr>
<tr>
<td>Y</td>
<td>Indicates the analyte was detected in both the sample and the associated method blank and was outside method blank acceptance criteria. This qualifier is reported for only values that are equal to or greater than the reporting limit.</td>
</tr>
<tr>
<td>ND</td>
<td>Reported value should not be used. Some or all of the quality control data for the analyte were outside acceptance criteria, and the presence or absence of the analyte cannot be determined from the data.</td>
</tr>
<tr>
<td>NR</td>
<td>Quality assurance/quality control of sample collection, handling, or processing is not sufficient to justify Illinois IEPA use of this result for Clean Water Act sections 305(b)/303(d) reporting and related purposes.</td>
</tr>
<tr>
<td>_A</td>
<td>The laboratory analysis was performed on an unpreserved or improperly preserved sample.</td>
</tr>
<tr>
<td>_P</td>
<td>Not detected. Note: For results reported by Illinois IEPA labs, if no Method Detection Limit is reported, then ND means not detected at the reported Reporting Limit.</td>
</tr>
<tr>
<td># (or &quot;z&quot;)</td>
<td>Not reported</td>
</tr>
<tr>
<td>* (or &quot;e&quot;)</td>
<td>Absent</td>
</tr>
</tbody>
</table>
APPENDIX B

METEOROLOGICAL VARIABLE EXTRACTION

The meteorological variables were extracted using Google Earth Engine Python API and `geemap` package.

```python
import os
import time
import geemap
import pandas as pd
import numpy as np
from datetime import datetime, timedelta
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from tqdm.notebook import tqdm

# Authenticates and initializes Earth Engine
import ee
try:
    ee.Initialize()
except Exception as e:
    ee.Authenticate()
    ee.Initialize()

os.chdir(r"I:\ALGAE")

df = pd.read_csv(r".\Data\EPA\2b_CHL_Monthly_with_stations.csv", index_col=0, parse_dates=['DATE_YM'])

def get_pixel_values(collection_id, start_date, end_date, lon, lat):
    # Convert datetime object to string
    start_date_str = start_date.strftime('%Y-%m-%d')
    end_date_str = end_date.strftime('%Y-%m-%d')

    try:
        ## Get the image collection
        collection = ee.ImageCollection(collection_id) \
        .filterDate(start_date_str, end_date_str) \
        .filterBounds(ee.Geometry.Point(lon, lat))

        # Get the mean of the image collection
        # It will return a single image
        image = collection.mean()

        # Get the band information at that given point
        band_dict = image.reduceRegion(reducer=ee.Reducer.mean(),
                                        geometry=ee.Geometry.Point(lon, lat),
                                        )
```
def create_band_df(collection_id, df):

    # Create an empty list to hold all the band values
    band_values = []
    for i in tqdm(range(df.shape[0])):
        band_info = get_pixel_values(collection_id=collection_id,
                                      start_date=df.iloc[i, 1],
                                      end_date=df.iloc[i, 1]+timedelta(days=30),
                                      lon=df.iloc[i, 3],
                                      lat=df.iloc[i, 2])
        band_values.append(band_info)

    # Combine all band values into a dataframe
    df_band_values = pd.DataFrame(band_values)

    # Join with the df
    df_join = df.join(df_band_values)
    # Drop the null values
    df_join = df_join.dropna()

    return df_join

# GRIDMET
df_GRIDMET = create_band_df("IDAHO_EPSCOR/GRIDMET", df)

df_GRIDMET.to_csv(r"\Data\EPA\_3b_CHL_Monthly_GRIDMET.csv")
APPENDIX C

LULC DATA EXTRACTION FROM NLCD

The LULC data was extracted from NLCD in Python using *rasterio* package.

```python
import os
import pandas as pd
import numpy as np
import rasterio
from rasterio.plot import show
from rasterio.windows import Window
import matplotlib.pyplot as plt
import geopandas as gpd
from rasterstats import zonal_stats

os.chdir(r"I:\ALGAE_VLMP")

def calcualte_zonal_stat(arr, condition, station_buffer, raster_affine):
    # Create an empty array with the shape of original array
    nlcd_class = np.zeros(arr.shape)
    # Change the values of that class to 1, otherwise its already 0
    nlcd_class[condition] = 1
    # Calculate zonal stat
    nlcd_class_zs = zonal_stats(station_buffer, nlcd_class, affine=raster_affine, stats=['sum'])
    nlcd_class_zs = [i['sum'] for i in nlcd_class_zs]
    del nlcd_class
    return nlcd_class_zs

def get_zonal_stat_as_table(nlcd_dir, buffer_dir, year):
    # Read nlcd raster
    with rasterio.open(nlcd_dir) as src:
        arr = src.read(1)
        raster_affine = src.transform
    # Read shp file
    station_buffer = gpd.read_file(buffer_dir)

    # Calculate all the zonal sum for all classes
    water = calcualte_zonal_stat(arr, (arr==11) | (arr==12), station_buffer, raster_affine)
    dev_open = calcualte_zonal_stat(arr, arr==21, station_buffer, raster_affine)
    dev_low = calcualte_zonal_stat(arr, arr==22, station_buffer, raster_affine)
    dev_med = calcualte_zonal_stat(arr, arr==23, station_buffer, raster_affine)
    dev_high = calcualte_zonal_stat(arr, arr==24, station_buffer, raster_affine)
    barren = calcualte_zonal_stat(arr, arr==31, station_buffer, raster_affine)
```

81
forest = calculate_zonal_stat(arr, (arr==41) | (arr==42) | (arr==43), station_buffer, raster_affine)
shrub = calculate_zonal_stat(arr, (arr==51) | (arr==52), station_buffer, raster_affine)
herb = calculate_zonal_stat(arr, arr==71, station_buffer, raster_affine)
crop = calculate_zonal_stat(arr, (arr==81) | (arr==82), station_buffer, raster_affine)
wet = calculate_zonal_stat(arr, (arr==90) | (arr==95), station_buffer, raster_affine)
impervious = calculate_zonal_stat(arr, (arr==22) | (arr==23) | (arr==24), station_buffer, raster_affine)

# Calculate the multiplication factor (to get percentage)
factor = 900.0 / (np.pi*np.square(3000))
# Here 900 is per pixel area in sq m (30m*30m), and the rest is area of circle (pi r squared)

# Update the geodataframe with all the class values
station_buffer['water'] = np.array(water) * factor
station_buffer['dev_open'] = np.array(dev_open) * factor
station_buffer['dev_low'] = np.array(dev_low) * factor
station_buffer['dev_med'] = np.array(dev_med) * factor
station_buffer['dev_high'] = np.array(dev_high) * factor
station_buffer['barren'] = np.array(barren) * factor
station_buffer['forest'] = np.array(forest) * factor
station_buffer['shrub'] = np.array(shrub) * factor
station_buffer['herb'] = np.array(herb) * factor
station_buffer['crop'] = np.array(crop) * factor
station_buffer['wet'] = np.array(wet) * factor
station_buffer['imperv'] = np.array(impervious) * factor

# Update the year
station_buffer['NLCD_year'] = year

# Remove the geometry column
station_buffer.drop(columns=['geometry'], inplace=True)

return station_buffer

nlcd_zs_2001 =
get_zonal_stat_as_table(nlcd_dir=r"\Data\NLCD\NLCD_2001_Proj.tif",
buffer_dir=r"\Data\Shapefiles\CHL_Data_Buff3km.shp",
year='2001')

nlcd_zs_2006 =
get_zonal_stat_as_table(nlcd_dir=r"\Data\NLCD\NLCD_2006_Proj.tif",
buffer_dir=r"\Data\Shapefiles\CHL_Data_Buff3km.shp",
year='2006')

nlcd_zs_2011 =
get_zonal_stat_as_table(nlcd_dir=r"\Data\NLCD\NLCD_2011_Proj.tif", 82
buffer_dir=r".\Data\Shapefiles\CHL_Data_Buff3km.shp",
    year='2011')

nlcd_zs_2016 =
get_zonal_stat_as_table(nlcd_dir=r".\Data\NLCD\NLCD_2016_Proj.tif",
    buffer_dir=r".\Data\Shapefiles\CHL_Data_Buff3km.shp",
    year='2016')

nlcd_join = pd.concat([nlcd_zs_2001, nlcd_zs_2006, nlcd_zs_2011,
    nlcd_zs_2016])

# Multiply all the values by 100 (converting scale 0-1 to 0-100)
    nlcd_join.iloc[:, 4:-1] = nlcd_join.iloc[:, 4:-1] * 100

nlcd_join.to_csv(r".\Data\3_NLCD_Class_Perc.csv")
APPENDIX D
LAKE MORPHOMETRY VARIABLE EXTRACTION

Lake morphometry variables were extracted using ArcGIS 10.4 Model Builder from the lake shapefiles and digital elevation model.
### APPENDIX E

#### DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES

<table>
<thead>
<tr>
<th>Type</th>
<th>Abr.</th>
<th>Variable Name</th>
<th>Unit</th>
<th>IEPA</th>
<th>VLMP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Meteorological</td>
<td>PRE</td>
<td>Precipitation</td>
<td>mm</td>
<td>0.25</td>
<td>13.78</td>
</tr>
<tr>
<td></td>
<td>RMX</td>
<td>Max relative humidity</td>
<td>%</td>
<td>67.18</td>
<td>99.76</td>
</tr>
<tr>
<td></td>
<td>RMN</td>
<td>Min relative humidity</td>
<td>%</td>
<td>21.94</td>
<td>69.30</td>
</tr>
<tr>
<td></td>
<td>SPH</td>
<td>Specific humidity</td>
<td>kg/kg</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>SRD</td>
<td>Shortwave radiation</td>
<td>W/m²</td>
<td>68.15</td>
<td>329.54</td>
</tr>
<tr>
<td></td>
<td>WDR</td>
<td>Wind direction</td>
<td>°N</td>
<td>117.84</td>
<td>253.33</td>
</tr>
<tr>
<td></td>
<td>TMN</td>
<td>Min temperature</td>
<td>K</td>
<td>262.62</td>
<td>296.39</td>
</tr>
<tr>
<td></td>
<td>TMX</td>
<td>Max temperature</td>
<td>K</td>
<td>271.94</td>
<td>308.80</td>
</tr>
<tr>
<td></td>
<td>WVS</td>
<td>Wind velocity</td>
<td>m/s</td>
<td>1.95</td>
<td>6.56</td>
</tr>
<tr>
<td>LULC</td>
<td>WAT</td>
<td>Water</td>
<td>%</td>
<td>0.00</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>FOR</td>
<td>Forest</td>
<td>%</td>
<td>0.00</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>CRP</td>
<td>Crop</td>
<td>%</td>
<td>0.00</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>WET</td>
<td>Wetland</td>
<td>%</td>
<td>0.00</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>IMP</td>
<td>Impervious Surface</td>
<td>%</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>Lake Morphometry</td>
<td>ARE</td>
<td>Area</td>
<td>km²</td>
<td>375.53</td>
<td>92.26</td>
</tr>
<tr>
<td></td>
<td>SHR</td>
<td>Shoreline</td>
<td>km</td>
<td>87.03</td>
<td>313.58</td>
</tr>
<tr>
<td></td>
<td>MXD</td>
<td>Max depth</td>
<td>m</td>
<td>0.00</td>
<td>21.82</td>
</tr>
<tr>
<td></td>
<td>MED</td>
<td>Mean depth</td>
<td>m</td>
<td>0.00</td>
<td>20.28</td>
</tr>
<tr>
<td></td>
<td>SDD</td>
<td>SD of depth</td>
<td>m</td>
<td>0.00</td>
<td>4.22</td>
</tr>
<tr>
<td></td>
<td>VOL</td>
<td>Volume</td>
<td>km³</td>
<td>0.00</td>
<td>6.09</td>
</tr>
<tr>
<td></td>
<td>IPQ</td>
<td>IsoPerimetic Inequality</td>
<td>/</td>
<td>0.01</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>DCM</td>
<td>Digital Compactness</td>
<td>/</td>
<td>0.04</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>PRE</td>
<td>Precipitation</td>
<td>mm</td>
<td>0.15</td>
<td>10.50</td>
</tr>
<tr>
<td></td>
<td>RMX</td>
<td>Max relative humidity</td>
<td>%</td>
<td>57.09</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>RMN</td>
<td>Min relative humidity</td>
<td>%</td>
<td>19.61</td>
<td>59.94</td>
</tr>
<tr>
<td></td>
<td>SPH</td>
<td>Specific humidity</td>
<td>kg/kg</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>SRD</td>
<td>Shortwave radiation</td>
<td>W/m²</td>
<td>75.10</td>
<td>335.23</td>
</tr>
<tr>
<td></td>
<td>WDR</td>
<td>Wind direction</td>
<td>°N</td>
<td>103.40</td>
<td>260.70</td>
</tr>
<tr>
<td></td>
<td>TMN</td>
<td>Min temperature</td>
<td>K</td>
<td>269.88</td>
<td>297.05</td>
</tr>
<tr>
<td></td>
<td>TMX</td>
<td>Max temperature</td>
<td>K</td>
<td>271.94</td>
<td>308.80</td>
</tr>
<tr>
<td></td>
<td>WVS</td>
<td>Wind velocity</td>
<td>m/s</td>
<td>2.23</td>
<td>6.04</td>
</tr>
<tr>
<td>LULC</td>
<td>WAT</td>
<td>Water</td>
<td>%</td>
<td>0.00</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>FOR</td>
<td>Forest</td>
<td>%</td>
<td>0.28</td>
<td>80.60</td>
</tr>
<tr>
<td></td>
<td>CRP</td>
<td>Crop</td>
<td>%</td>
<td>0.05</td>
<td>82.48</td>
</tr>
<tr>
<td></td>
<td>WET</td>
<td>Wetland</td>
<td>%</td>
<td>0.00</td>
<td>22.97</td>
</tr>
<tr>
<td></td>
<td>IMP</td>
<td>Impervious Surface</td>
<td>%</td>
<td>0.34</td>
<td>85.69</td>
</tr>
<tr>
<td>Lake Morphometry</td>
<td>ARE</td>
<td>Area</td>
<td>km²</td>
<td>2427.3</td>
<td>15.79</td>
</tr>
<tr>
<td></td>
<td>SHR</td>
<td>Shoreline</td>
<td>km</td>
<td>227.83</td>
<td>112.94</td>
</tr>
<tr>
<td></td>
<td>MXD</td>
<td>Max depth</td>
<td>m</td>
<td>0.00</td>
<td>20.32</td>
</tr>
<tr>
<td></td>
<td>MED</td>
<td>Mean depth</td>
<td>m</td>
<td>0.00</td>
<td>20.28</td>
</tr>
<tr>
<td></td>
<td>SDD</td>
<td>SD of depth</td>
<td>m</td>
<td>0.00</td>
<td>1.78</td>
</tr>
<tr>
<td></td>
<td>VOL</td>
<td>Volume</td>
<td>km³</td>
<td>0.00</td>
<td>22066</td>
</tr>
<tr>
<td></td>
<td>IPQ</td>
<td>IsoPerimetic Inequality</td>
<td>/</td>
<td>0.01</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>DCM</td>
<td>Digital Compactness</td>
<td>/</td>
<td>0.04</td>
<td>0.74</td>
</tr>
</tbody>
</table>

85
VITA

Graduate School
Southern Illinois University

Supria Sarkar
supriabarsha@gmail.com; supria.sarkar@siu.edu

University of Dhaka, Bangladesh
Master of Science, Environment Science, May 2018

University of Dhaka, Bangladesh
Bachelor of Science, Soil Water and Environment, November 2016

Special Honors and Awards:
    Bangladesh Sweden Trust Fund Travel Scholarship 2021

Thesis Paper Title:
    Using Machine Learning to Understand the Spatiotemporal Variability of Harmful Algae Blooms in Illinois Waters

Major Professor: Dr. Ruopu Li