5-23-2023

# Remote Sensing and Artificial Intelligence-Based Modeling and Prediction of Harmful Algal Blooms in Lake Pontchartrain

Ian Smith
*Louisiana State University and Agricultural and Mechanical College*

# REMOTE SENSING AND ARTIFICIAL INTELLIGENCE-BASED MODELING AND PREDICTION OF HARMFUL ALGAL BLOOMS IN LAKE PONTCHARTRAIN

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Civil and Environmental Engineering

by
Ian Mathew Smith
B.S., Louisiana State University, 2019
August 2023

And now for something completely different.

—Monty Python
*Flying Circus*

## Acknowledgments

First and foremost I would like to thank God, as it is only by his design and grace that I was able to write this thesis. I would like to thank my parents for their love and support, especially for the financial support this final semester. Without it, I would not have been in a position to finish my research this semester, and most likely would not have been able to graduate. I would like to acknowledge Dr. Zhiqiang Deng, my graduate advisor, for guiding me and helping write this thesis. Without his guidance and teachings, producing research results that were both a scientific advancement as well as being scientifically competent would not have been possible. I would also like to thank Dr. Samuel Snow, while he was not directly involved in the production of this thesis, his involvement in my life as my undergraduate advisor put me in the position to succeed as a graduate student. To my friends and fellow graduate students, Xiangjie Wang and Saber Aradpour, thank you from the bottom of my heart for always helping me when I became idle in my research, and for giving me guidance in areas in which I lack expertise. To one of my best friends, Pujan Shrestha, thank you for always pushing me to not give up, helping me with LaTeX coding, for proof reading, and for sitting through my practice presentations when you had much better things to do. Finally I would like to thank myself, for believing in myself, and for not quitting when things looked impossible.

## Table of Contents

## List of Tables

## List of Figures

## Abstract

Harmful Algal Blooms (HABs) and particularly toxic cyanobacterial harmful algal blooms (CyanoHABs) have become a growing threat to the environment, economy, communities, and human and animal health. This is particularly true for Lake Pontchartrain. Forecasting the occurrence of cyanobacterial HABs in Lake Pontchartrain is a process that currently does not exist, as nowcasting by the NCCOS Algal Bloom Monitoring System or the similar system implemented by the U.S. EPA is currently the only method to monitor HAB production in the Lake. This thesis made this process possible by identifying antecedent environmental conditions controlling CyanoHABs, describing the conditions using NASA satellite remote sensing data from the MODIS-Aqua, and finally simulating the conditions and associated CyanoHABs by developing eight forecasting models with the lead-time of 15-22 days for predicting NCCOS Clycano index value representing the level of CyanoHABs. Specifically, eight Random Forest models were created by using the WEKA platform and two years of time series data from 2021 – 2022 for NCCOS Clycano index values and corresponding satellite remote sensing data for chlorophyll-a concentration, sea surface temperature and reflectance bands. Additionally, the models were validated with data from 2019. Model forecasting results based on the training data indicate that all eight models are capable of forecasting Clycano index values with a very high correlation coefficient of 0.90 or higher and MAE of 10 and RMSE of 20 or lower. The theoretical significance of using the eight forecasting models is that the negative impact of cloud cover on the availability of remote sensing data can be minimized, greatly expanding the application of satellite remote sensing data. The practical significance of the eight forecasting models is that they make it possible to forecast CyanoHABs on a daily basis

and thereby inform water quality programs of where and when CyanoHABs are likely to occur so that managers can proactively respond to CyanoHAB events, greatly reducing the CyanoHAB risk to the public health.

# Chapter 1. Introduction

Photosynthesizing organisms such as planktonic micro and macro algae are the foundation of almost all trophic interactions within the aquatic photic layer [5]. Although these organisms are an important component of primary production in the food web, when large concentrations of these organisms form and die they can cause oxygen depletion, hypoxia/anoxia, by decomposition [28]. In addition, approximately 300 species are known to produce toxins that can negatively affect the fauna and ecosystem [35]. HABs are becoming more common in conjunction with climate change raising temperatures [28].

The microalgae that form HABs are diverse and belong to six different groups that include: diatoms, dinoflagellates, haptophytes, raphidophytes, cyanophytes, and pelagophytes [22]. The dinoflagellate phylum is the main contributing group that form hypoxic HABs; however, the cyanophyte phylum is a major contributor associated with toxic blooms [35]. As briefly mentioned above, HABs can cause hypoxia, anoxia, and synthesize toxins. The toxins that are produced have various effects by either entering the food web tropically or coming into contact physically with fauna [35]. A notable aspect is problems caused by the toxins are not usually associated with high biomass blooms, as toxic events can occur from very low concentrations of the suspect algae [22]. The most prevalent harmful effects on human health caused by toxic blooms are: the consumption of filter feeding shellfish, consuming the fish that bioaccumulate the toxins from feeding on the shellfish, respiratory problems from aerosols released by the toxic seawater, and skin irritations caused by contact with the skin [22]. The local communities that border Lake Pontchartrain including Mandeville, Metairie and New Orleans, as well as individuals that travel to these communities for recreation can be impacted by these health risks. The

Figures 1.1 & 1.2 below, show a true color image and an aerial photo of a bloom in the northern part of the lake near Mandeville. While large biomass blooms can produce toxins, their main detrimental effects are environmental issues such as anoxia and food-web changes as well as economic issues associated with recreational and commercial waters including foul odor and congested water that makes swimming difficult [22]. This thesis will focus upon detecting these high biomass blooms.

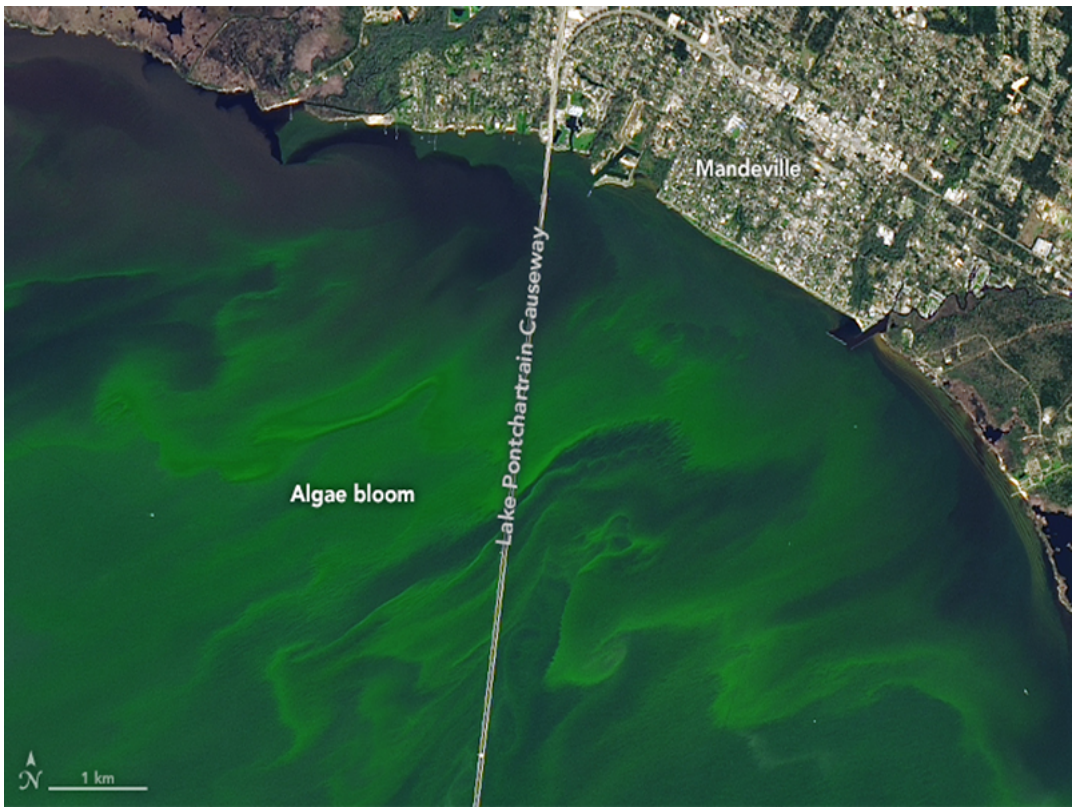


Figure 1.1. An algae bloom near Mandeville, LA (photo courtesy of Tulane University)

Hypoxia and anoxia pose the greatest threat to aquatic fauna as sometimes thousands suffocate due to the inability to leave the hypoxic zone [40]. A fully anoxic condition is far less likely to occur in a large lake or open ocean. However, when large enough areas are hypoxic, it is enough to kill large organisms that need a significant amount of

oxygen and/or small organisms that cannot escape the large areas which are sometimes several 100 square kilometers [28]. When considering wind, currents and rivers many hypoxic events impact ecosystems far from where the source bloom originated [15]. Lake Pontchartrain provides spawning grounds for many species of sport fish such as speckled trout, redfish, and drum as well as supporting a shellfish and crustacean population [1]. Oysters reefs are not as prevalent throughout the lake as they require water with greater salinity, and most are located near the eastern most side of the lake [1]. However, Lake Pontchartrain does affect oyster populations as it can discharge large amounts of freshwater into Lake Borgne and subsequent surrounding marsh bays, especially when the Bonne Carré spillway is opened [1].

Figure 1.2. An Aerial Photo of an Algae bloom in Lake Pontchartrain, LA (photo courtesy of The Lake Pontchartrain Foundation)

Interestingly enough a main cause of the problem, light, may also be the source of a solution to help predict these harmful blooms. The advancement of satellites in the 1980's led to the usage of surface reflected light to measure certain constituents on a large spatial scale [33]. Machine learning algorithms and computing software has also drastically improved in the past 20 years, and is an exciting field of science and engineering that shows promise in a multitude of fields. The question is: can these technologies when coupled together use free and openly available online data to predict CyanoHABs occurring in large bodies of water where collecting consistent wide spread in-situ data for these regions is expensive and difficult?

The objective of this thesis was to achieve this goal by first collecting the data and second training an algorithm that can forecast the occurrence of harmful algae blooms. Even though these sources are free, having a computer is necessary in order to utilize them. The data for this thesis utilizes satellite remote sensing data from two different satellites but must be processed from its raw form. Data processing programs such as SeaDas, Matlab (or equivalent coding program such as Python or R studio), and Excel are needed to manipulate the raw data. A platform able to use a machine learning algorithm is needed to be able to utilize the raw data. This thesis used the machine learning platform WEKA, which is a user friendly platform that makes using machine learning algorithms easy to use. The overall flow chart of this thesis is shown by the figure below.

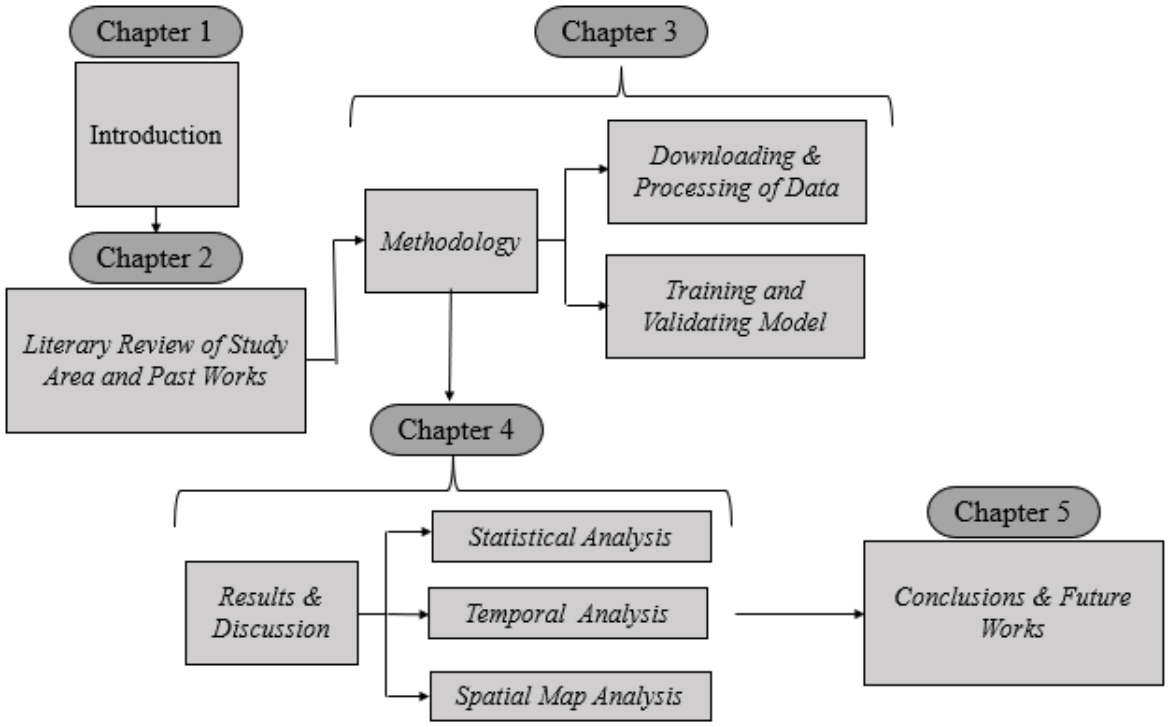


Figure 1.3. Thesis Flow Chart

## Chapter 2. Literature Review

The basic factors that control algal growth are water temperature, salinity, nutrients (nitrogen (N) and phosphorous(P)), and light. Although these factors are essential, many other variables such as natural physical processes, anthropogenic processes, and organism behavior are responsible for where and when HABs form [15]. The cause of increased HAB occurrence in the past few decades has mainly been blamed on anthropogenic effects, and dependent on hydrographic and ecological location [12, 15, 22, 28, 35, 40]. The main anthropogenic effect, nutrient loading, is caused by the increased use of synthetic fertilizers which have tripled the export of organic phosphorus and nitrogen to coastal regions [12, 15, 22, 28, 40]. Population growth is a main contributor as well, as it can drastically alter the landscape contributing to increase land runoff as well as an increase in large sewage inputs [15, 22]. A prime and early example of anthropogenic effects effecting HAB occurrence is supported by a study in the Seto Inland Sea region of Japan where population and industrial development from 1968–1976 showed a 6-fold increase in HABs each year, with N and P concentrations increasing 30- and 5-fold respectively [37]. Local anthropogenic effects are not as big of an issue as their origin can be pinpointed, whereas non-point sources of nutrient loading, such as atmospheric deposition of nitrogen, are of much larger concern as their source is hard to identify [22]. These non-point anthropogenic sources couple with non-point physical processes such as wind, rivers, tide and ocean currents are all major causes of HABs forming in areas where the source is not local [15, 22].

## 2.1. Study Location

This thesis's study location is Lake Pontchartrain, which is a large estuary that receives continuous and episodic freshwater inputs [24, 27, 29, 39]. The lake naturally receives runoff from the city of New Orleans (which it borders) and receives nutrients and fresh water from several northern tributaries and the fresher Lake Maurepas [29]. Salt water is introduced into Lake Pontchartrain three main ways: the two natural tidal passes connecting it to Lake Borgne and the Inner Harbor Navigation Canal (IHNC) which is connected to the Mississippi River Gulf Outlet (MRGO) [24, 27]. When flooding of New Orleans from the Mississippi River is of concern, the Bonne Carré Spillway is opened allowing river water to enter the lake [10, 24, 27, 29, 36]. The Spillway has been opened numerous times with the most recent being in 2020, twice in 2019, 2018, 2016, and 2011, each time a high biomass bloom has been recorded in the lake [10, 32, 37]. During the years that the spillway is opened it is almost always expected that a large toxic bloom will occur [29]. The driving limiting environmental factors that produce toxic HABs in Lake Pontchartrain are linked to salinity and nutrient loading [21, 24]. Salinity is an interesting factor as many toxic blooms belonging to cyanophytes thrive in a saltier environment, but phytoplankton and many other algae species do not [21]. For this reason, it was believed that closing the MRGO in 2009 and reducing the amount of salt water introduced to the lake would help decrease the amount of toxic HABs. Studies have shown that since its closure Lake Pontchartrain's mean salinity has decreased [1, 24]. However, according to the NCCOS's harmful algal bloom monitoring system, during the year 2021 when the Bonne Carré spillway was not opened, there was a significant increase in the number of

high biomass blooms as compared to previous years. This is of concern and the driving force behind this thesis; because, if this trend continues predicting these blooms in the lake cannot rely solely on considering anthropogenic events such as opening the spillway.

Nutrient loading is the other main driving factor as many cyanophytes rely on dinitrogen fixation for growth [21]. Lake Pontchartrain's primary production is normally N limited; however, during large nutrient loading events it has been known to switch to P limiting [27]. This is important for Lake Pontchartrain as physical resuspension of soluble reactive phosphorous (SRP) can be dominated by wind waves in shallow lakes [27]. Even though the lake is the second largest saltwater lake in the U.S. with a surface area of 1630 km2, the average depth of the Lake is only 3.5 to 4 meters [10, 29]. Due to this large surface area and shallow depth accompanied with weak tidal influence the lake is unstratified and circulation is primarily wind driven [10].

### 2.2. Past Works

This next section will be devoted to discussing past works published regarding the prediction of CyanoHABs with the application of remote sensing. In the past predicting CyanoHABs has mainly been achieved by using in-situ empirical data driven models such as environmental fluid dynamics code (EFDC) coupled with machine learning algorithms to predict Chlorophyll-a concentrations (which is a good indicator for primary production) to then predict if an CyanoHAB will or will not occur [11, 19, 20, 23, 36, 38, 39]. Over the years, many machine learning algorithms have been developed for use in this endeavor. They include: artificial neural network (ANN) also known as multiple layer perceptron (MLP), back-propagation (BP) neural network, generalized regression neural network

(GRNN), convolutional neural network (CNN), long short-term memory (LSTM), support vector machine (SVM), support vector regression (SVR), K-nearest neighbor (KNN), adaptive boosting (AdaBoost), gradient boosting decision tree (GBDT), extreme gradient boosted decision trees (XGBoost), and random forest (RF). Of these algorithms MLP, GBDT, XGBoost, AdaBoost, and RF models stand out as the most effective when dealing with complex nonlinear phenomena.

MLP or ANN is a widely used machine learning algorithm and it functions by simulating the human brain by building a multilayer network to achieve the goal of prediction [38]. MLP is usually constructed with three different layers consisting of an input layer, a hidden layer, and an output layer [38]. The input layer is just as it sounds and consists of the input categories as nodes which feed the hidden layers. There can be any number of hidden layers usually with the same or less amount of input nodes effectually creating a nodal network that uses sigmoid and ReLU functions to predict the final output layer [38].

AdaBoost, GBDT, XGBoost are what is known as boosting algorithms which focuses on generating a significantly accurate prediction by combining moderately inaccurate or weak predictors [13]. The AdaBoost algorithm is a very common algorithm that has many different sub algorithms that deal with either classification problems or regression problems [30]. For this thesis the additive regression model of AdaBoost will be considered an option. AdaBoost's main goal is that it garners more attention to samples that predicted wrong values [38]. It does this so that in the next iteration of training they are more likely to make a correct prediction [38]. AdaBoost.R is an algorithm created by Freude and Schapire to solve regression problems by extending their AdaBoost.M2 algorithm [30]. This AdaBoost.R algorithm is the base framework of many other AdaBoost

algorithms that try and improve on AdaBoost.R's flaws [27, 38]. The major flaw with AdaBoost that is studied the most is it's loss function, which is incorporated to measure the machine's performance, changes from iteration to iteration [30]. A major strength of AdaBoost and boosting in general is that there are many different weak learners or base learners to choose from, although many are tree-based learners [2]. GBDT is very similar to AdaBoost with the difference being that its goal is to reduce the loss for each iteration [14, 38]. This is achieved by fitting the approximation of the loss using a negative gradient of loss function. When gradient boosting regression trees it is found that the results are robust and highly interpretable as well as competitive in their results especially when using noisy data [14]. A strong aspect of GDBT is it's ability to obtain accurate feature importance by averaging the importance in each single decision tree, which leads to a better understanding of the data [38]. XGBoost, an improved gradient boosting algorithm, is a scalable end-to-end tree boosting ensemble [9]. It is particularly powerful as it can use out-of-core computation and parallel processes that allows for very large data computation faster than other boosting programs such as AdaBoost [9].

RFs are very similar to boosting algorithms as they are both ensemble learners. RF's use a combination of learners, in this case the decision trees, to formulate a more accurate prediction [2, 26]. A RF model grows an ensemble of decision trees where each tree is grown by a random selection from the data given in a boot-strapped training set [6]. The RF model aggregates the votes from the decision trees to make a prediction, which is called bagging [7]. The boot-strapped dataset is created when data from the original dataset is randomly selected and a new dataset in a different order but same size is created [6, 7]. Along with being in a different order the data may be chosen more than once

which leaves data left over called out-of-bag dataset [6, 7]. The out-of-bag dataset is usually about 30 percent of the original dataset, and it is used to test the predictions made by the decision trees and evaluate their performance [6, 7]. Regression trees are similar to decision trees but instead of having two leaves built off one node with a yes or no answer [6]. Instead there are multiple nodes and leaves built upon the base node and an optimal numerical value is found [6]. Along with boosting, RF models are especially effective at solving non-linear regression models as they can isolate irregular correlations in a data set.

It can be seen that using in-situ data points with machine learning algorithms have been proven to work; however, maintaining water quality sensors and performing laboratory tests for continuous temporal and spatial data can prove to be very expensive [39]. Thus, while initially costly to launch satellites into space, the long term reduced costs of using remote sensing satellite data for HAB monitoring and prediction is appealing. This idea was first introduced by Steidiner and Haddad in 1981 by their work using the onboard sensor equipped to the Nimbus-7 [35], and influenced the launch of SeaWiFS, MERIS, MODIS and Sentinel satellites [16]. These satellite programs have proved fruitful. In 2011 a study using machine learning algorithms RF and SVM was conducted using only MODIS and MERIS remote sensing data to determine where and when to sample viable in-situ data [31]. In a recent study, Hill, P.R. et al predicted HABs off of Florida's coast line using an iterative combination of first using CNNs followed by LSTMs models finishing with either a SVM, MLP, or RF model then comparing the models for accuracy. A year later a newer study for the same area by Izadi M. et al was very similar to Hill's with the main difference being that the study considers the lag time between occurrence of a bloom and the time it takes for the variable to have a maximum impact on the bloom

propagation as a major point of emphasis. Izadi also used in-situ concentration data for the specific algal species Karenia brevis as a dependent variable coupled with MODIS-Aqua data as an independent variable, and did not use the iterative combination of models as Hill, as only RF, SVM and XGBoost were used and compared.

To differ from these studies and many others like it this, thesis will address a different study area, Lake Pontchartrain, and will use MODIS-Aqua remote sensing data and NCCOS Sentinel 3 Clycano index value (CIV) data to train a model without in-situ chlorophyll or algal species concentration. Like Izadi's study, lag time will be a large focus of this study but will differ in one aspect by using open-source data such as MODIS-Aqua ocean color data in set increments ahead of CIV data.

# Chapter 3. Methodology

To begin the development of the forecasting model, raw data comprised of MODIS-Aqua satellite ocean color data and NCCOS Sentential 3 Clyano index value needed to be obtained and processed into a form usable by the machine learning algorithm. During the year 2021 there were more CyanoHAB events than previous years making 2021 a good training set as it has the most CIV data. Level 2 satellite ocean color data as well as sea surface temperature (SST) in Celsius obtained from the MODIS-Aqua Satellite were downloaded from the NASA's ocean color website ranging from January to December 2021. It is important for the model to be able to predict which days are going to have a bloom, but it is just as important that the model predict days that do not have a bloom. Algae blooms being a seasonal occurrence, it is important that the model can predict days that do not have a bloom when blooms are more likely to occur i.e., in the summer months between July and August. Surprisingly only a year after the high occurrence of blooms in 2021, in 2022 there were very few large magnitude blooms in lake Pontchartrain, especially in the summer months when historically they occur the most. For this reason, 2022 remote sensing data from January to December was also downloaded and included in the model's training dataset.

## 3.1. Variable Description and Importance to CyanoHAB Detection

The ocean color data (OC) includes the variables: sea surface reflectance (Rrs) measured as ($sr^{-1}$) at the ten different wavelengths of 412nm, 443nm, 469nm, 488nm, 531nm, 547nm, 555nm, 645nm, 667nm, and 667nm, the attenuation coefficient (Kd) measured as ($m^{-1}$) at 490 nm, chlorophyll concentration (chlor-a) measured as (mg $m^{-3}$),

particulate organic carbon concentration (poc) measured as (mg $m^{-3}$), particulate inorganic carbon (pic) measured as (mol $m^{-3}$), instantaneous photosynthetically available radiation (ipar) measured as (Einstein $m^{-2}s^{-1}$), photosynthetically available radiation (par) measured as (Einstein $m^{-2}d^{-1}$), normalized florescence height (nflh) measured as (mW $cm^{-2}$ $\mu m^{-1}$ $sr^{-1}$), the dimensionless aerosol optical thickness at 869nm (aot_869), and the dimensionless aerosol angstrom exponent (angstrom).

The Rrs terms make up the bulk of the data and are the backbone of almost all remote sensing data algorithms. Rrs terms are derived from the corrected downward radiance (Ed) and water leaving irradiance term (Lw) and is expressed for a specific wavelength $\lambda$ as: [18]

$$Rrs(\lambda) = Lw(\lambda)/Ed(\lambda) \tag{3.1}$$

This equation is a decimal ratio of the amount of light being reflected or scattered back up out of the water column at a certain wave length [18]. The above equation is a simplified equation, as the Ed term is far more complex and uses terms that correct for atmospheric conditions, but the simplified equation suffices for the knowledge needed to understand this study. Figure (3.1) below is a great diagram visually explaining the way that the sensors detect light.

The terms aot_869, and angstrom are by-products from deriving the Rrs terms and are a diagnostic of the Rrs algorithms performance [4]. These bands are important parameters as they can detect when the Rrs bands need to be corrected for atmospheric conditions such as high ozone levels, cloud cover, and general smog [4, 18].

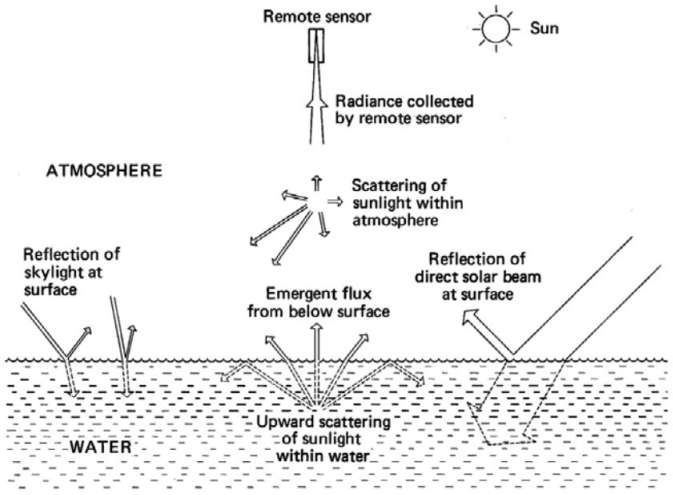


Figure 3.1. A diagram of the origins of light detected by a remote sensing device above a body of water. [18]

Rrs bands are very useful in estimating the constituents in the water column as certain organic and inorganic substances absorb or scatter certain wave lengths of light [18]. For example chlorophyll absorbs light in the blue and red nm range and reflects in the green range; therefore, theoretically in the presence of chlorophyll the sea surface reflectance ratio in the green wave length would be high and low in the blue and red [18]. For this reason many empirical equations are used referencing different Rrs band ratios to predict the chlorophyll- a concentration [18, 4]. The chlorophyll concentrations used in this thesis's model is the chlorophyll-a pigment. Chlorophyll has three common pigments (chlorophyll-a, -b, and -c) and all can be found in HABs; however, chlorophyll-a has the strongest correlation for measuring algal growth in aquatic environments [17].For this reason chlorophyll-a measurements from the MODIS-Aqua OC data is a very important variable to consider in the predictive model. The chlorophyll-a concentration is determined by

a combination of the CI equation and the OCx equations they are as follows [4].

$$CI = R_{rs}(\lambda_{green}) - \left[R_{rs}(\lambda_{blue}) + \frac{\lambda_{green} - \lambda_{blue}}{\lambda_{red} - \lambda_{blue}} \times (R_{rs}(\lambda_{red}) - R_{rs}(\lambda_{blue}))\right] \tag{3.2}$$

which is then inserted into the formula below with the coeffiecients $a_{0CL}$=-0.4287 & $a_{1CL}$= 230.47

$$chlor\_a_{CI} = 10(a_{0CI} + a_{1CI} \times CI) \tag{3.3}$$

The OCx formula [4] is next shown as

$$\log_{10}(chlor\_a) = a_0 + \sum_{i=1}^{4} a_i \left(\log_{10}\left(\frac{R_{rs}(\lambda_{blue})}{R_{rs}(\lambda_{green})}\right)\right)^i \tag{3.4}$$

The formula has several coefficients that are determined by which specific sensor is using them. A chart is shown below represented by Table 3.1

Table 3.1. Three popular remote sensing satellite's Rrs bands and coefficients for the chlor-a OCx algorithm. [4]

| sensor | Algorithm | OCx Rrs used(blue/green) | a(0,1,2,3,4) |
|---|---|---|---|
| SeaWiFis | OC4, CI | Rrs(433 489 510)/Rrs555 | 0.32814; -3.20725; 3.22969; -1.36769; -0.81739 |
| MODIS | OC3M, CI | Rrs(443 488)/Rrs547 | .26294; -2.64669; 1.28364; 1.08209;-1.76828 |
| MERIS | OCE, CI | Rrs(443 489 510)/Rrs555 | 0.42487; -3.20974; 2.89721; -0.75258; -0.98259 |

For chlorophyll concentrations below 0.25 mg $m^{-3}$ the CI algorithm is used and for chlorophyll concentrations above 0.35 mg $m^{-3}$ the OCx algorithm is used. The in-between values are calculated by a third algorithm [4]

$$chlor\_a = \frac{chlor\_a_{CI}(t_2 - chlor\_a_{CI})}{t_2 - t_1} + \frac{chlor\_a_{OCx}(chlor\_a_{CI} - t_1)}{t_2 - t_1} \tag{3.5}$$

with $t_1 = 0.25$ and $t_2 = 0.35$

The variable nflh stands for normalized fluorescence line height and is very closely related to chlorophyll as it is the relative measure of the water-leaving radiance (Lw) as-

sociated with chlorophyll florescence [4]. It is calculated as the difference between the observed nLw(678) and the linearly interpolated nLw(678) from two adjacent surrounding bands. When algae or phytoplankton undergo photosynthesis the light energy absorbed to undergo this process can be re-emitted as fluorescence [18]. This re-emittance is measurable and is a good indicator of the presence of chlorophyll and primary production in the water column [18]. As another indicator of chlorophyll it has a high chance of being a useful predictive variable in the HAB model. Its generic algorithm is as follows [4]

$$nflh = nLw(678) - (\frac{70}{81}) \times nLw(667) - (\frac{11}{81}) \times nLw(748) \tag{3.6}$$

where the nLw term is the normalized water leaving radiance at the specified wavelength

The Kd_490nm term, is known as the diffusive attenuation coefficient of downwelling irradiance at 490 nm which is one of the most important optical properties of ocean water, especially when considering coastal turbid waters as it is an indicator of turbidity and possible presence of nutrients in the eutrophic zone[17, 8, 18]. The Kd_490 term, like chlor-a variable, uses an equation that references Rrs bands that coincide with the sensor being used. The equation for Kd_490 is calculated in two parts and is as follows

$$\log_{10}(K_{bio}(490)) = a_0 + \sum_{i=1}^{4} a_i \left( \log_{10}(\frac{R_{rs}\lambda_{blue}}{R_{rs}\lambda_{green}}) \right)^i \tag{3.7}$$

the equation is solved for $K_{bio}$ as a fourth order polynomial then inserted into

$$Kd\_490 = K_{bio}(490) + 0.0166 \tag{3.8}$$

The coefficients $a_0$ & $a_i$ as well as the two Rrs bands for equation 3.7 are determined from the below table

Table 3.2. Three popular remote sensing satellite's Rrs blue and green bands and coefficients for the $K_{bio}$ algorithm. [4]

| | sensor | blue | green | a0 | a1 | a2 | a3 | a4 |
|---|---|---|---|---|---|---|---|---|
| KD2S | SeaWiFis | 490 | 555 | -.8515 | -1.8263 | 1.8714 | -2.4414 | -1.0690 |
| KD2M | MODIS | 488 | 547 | -.08813 | -2.0584 | 2.5878 | -3.4885 | -1.5061 |
| KD2E | MERIS | 490 | 560 | -0.8641 | -1.6549 | 2.0112 | -2.5174 | -1.1035 |

In a study on Kd_490nm it was found that the variable was a proxy for the growth of phytoplankton, and in another study cited by Izadi et al it was observed that there was a high correlation between chlorophyll-a and Kd_490 during red tide events in the Persian Gulf [8]. Considering Lake Pontchartrain is a very turbid lake, the Kd_490 variable is an important variable to include within the predictive model of HABs. For optically complex waters, such as Lake Pontchartrain, scattering due to suspended solids, chlor-a, and CDOM (colored dissovled organic matter) is high; therefore, a reference ratio in the longer wave lengths, such as the red nm range, for the Kd_490 variable is needed [3]. Due to this phenomenon the importance of the Kd_490 variable in the MODIS-Aqua OC data is likely to fail to truly capture the true effective attenuation coefficient for Lake Pontchartrain as it is not corrected with a red band Rrs ratio. However, it will still a be a good judge of when the waters are more optically clear which would be an indicator of a clearer photic zone, and subsequently a higher change of an algal bloom .

The next important variables are the PAR and iPAR variables which stand for photosynthetically available radiation and instantaneous photosynthetically available radiation respectfully [4]. PAR is the measurement of the daily average photosynthetically available radiation at the ocean's surface and is defined as the quantum energy flux from the sun in the visible wavelength spectrum (400-700nm) [4]. The PAR measurement is an indica-

tor of primary production as the wavelengths within its range are used by phytoplankton for photosynthesis and can be a determining factor on what kind of algae species forms and its overall distribution worldwide [8, 4]. Unlike PAR which is a daily average, iPAR is the photosynthetically availbe radiation at the imediate time stamp of when the remote sensing data was taken [8, 4]. The iPAR variable is in fact a companion product to the aforementioned nflh variable and it can be used in combination to estimate the fluorescence quantum yield [4]. MODIS offers two daytime iPAR readings, one from the Terra satellite and the other from the Aqua satellite, and the iPAR reading in this study uses the ipar from the Aqua satellite. The PAR reading is a daily reading of iPAR, and since Terra crosses the equator approximately three hours before Aqua the total MODIS PAR is added to the Aqua data as a daily average of those two readings [34]. PAR and iPAR are indicators of primary production, and it makes this variable extremely pertinent to the development of the predictive model. The equation for iPAR is found below [4]

$$iPAR = \frac{1}{hc} \int_{700}^{400} \lambda E_d(\lambda, 0-) d\lambda \tag{3.9}$$

- h = Planck's constant
- c = Speed of light
- $E_d(\lambda,0-)$ = downwelling irradiance just below the sea surface

Along with the aforementioned OC variables, sea surface temperature is also a critical variable when considering algal blooms and is also collected from the MODIS-Aqua Satellite, but by a separate sensor. For algal growth, SST plays a critical role in the timing of algal blooms, and it is a very important parameter to consider in the predictive

model. [27, 11, 15, 16, 17, 18, 20, 22, 28, 37, 40]. The SST MODIS algorithm returns what is called the skin surface temperature using the long-wave infrared (LWIR) spectral bands 11 $\mu m$ and 12 $\mu m$ [4]. It is commonly referred to as skin temperature as the infrared originates from the surface thermal skin layer of the ocean, not the underlying ocean water [4]. This skin layer is less than 1mm thick and by rule this skin layer is cooler than the underlying water and typically the relationship this skin temperature and the subsurface is rather stable when wind speed is above 6 m/s [4]. To accommodate for when the conditions are unstable, coefficients were developed for the algorithm to correct them [4]. The current SST algorithm is a modified version of a nonlinear algorithm that utilizes empirical coefficients that are derived by regression of collocated in-situ and satellite measurements for a distinct location based upon a latitude zone and the month of year the data falls in [4]. The generic algorithm for SST in Celsius is as follows [4]

$$SST = a_{ij0} + a_{ij1} B_{T11}\mu m + a_{ij2}(B_{T11}\mu m - B_{T12}\mu m) T_{sfc}$$

$$+ \mathrm{a}_{ij3} \sec(\theta - 1)(B_{T11}\mu m - B_{T12}\mu m) \quad + a_{ij4}(mirror) + a_{ij5}(\theta^*) + a_{ij6}(\theta^2) \quad (3.10)$$

- $\mathrm{BT}_{11\mu m}$ = Brightness Temperature(BT in the $11\mu m$ channel
- $\mathrm{BT}_{12\mu m}$ = Brightness Temperature(BT in the $12\mu m$ channel
- $\mathrm{T}_{\mathrm{sfc}}$ = Reference SST
- $\theta$ = sensor zenith angle
- $\theta^*$ = sensor zenith angle is made negative for pixels in the first half of the scan line
- mirror = mirror side number (0 or 1)
- coefficients aij = algorithm coefficient set for month of year, i, and latitude zone, j

The final two variables that were downloaded were particulate organic carbon (POC) and particulate inorganic carbon (PIC). After compiling the data it was noticed that the PIC variable had a considerable amount of data that was labeled as N/A, and when it did have viable data all the other data was not viable. For this reason it was decided to eliminate the variable entirely. Fortunately, the POC variable was readily available and was considered in the model. As it can be inferred the POC algorithm records the concentration of particulate organic carbon in mg $m^{-3}$ in the water column. It does this by using blue to green Rrs band ratios and empirical relationships from in-situ measurements [4]. The ability to be able to sense organic material in the water column can be a vital constituent in developing a model that can predict HABs since HABs are made of organic material. Certain levels of POC concentrations can also be used to detect the prerequisites needed for HABs to form as there would be certain levels that indicate a viable nutrient level [24, 18]. The POC algorithm is as follows below [4]

$$poc = a \times (\frac{Rrs(443)}{Rrs(555)})^b \tag{3.11}$$

where a = 203.2 and b = -1.034 and in cases where Rrs(555)is not available, an empirical relationship is used to give an equivalence relating available bands; for MODIS the Rrs bands used are 443, and 547 [4].

### 3.2. Pre-Processing and Limitations of Data

Previous bloom event data was obtained from NCCOS's harmful algal bloom monitoring system which uses the Sentinel 3a and 3b satellites to generate an index value, called the Clycano index value (CIV), measuring the intensity of the bloom. This data was obtained as a geo-tiff that is labeled as coming from either the Sentinel a or b

satellites. Both CIV and MODIS-Aqua ocean color data were imported and viewed into SeaDAS v 8.3 for processing. Before downloading, the OC data must be re-projected into the correct coordinate plane. The coordinate plane chosen was the World Geodetic System 1984 UTM North Hemisphere Zone 15 and all data with a corresponding latitude and longitude was downloaded and processed in this coordinate system. It is important to note that MODIS-Aqua satellite data is not available every single day, and that is a huge limitation in training the model as well as validating it. Figures (3.2 & 3.3) are a good example of viable day for MODIS-Aqua OC data. Figures (3.4 & 3.5) show, respectively, examples of what non-viable and viable days look like for CIV data. Initial factors such as swath coverage affect the availability of data, and viable OC data needs to have matching viable SST data to accompany it. Furthermore, even if the data is available there are four SeaDAS masks, LAND, HIGLINT, STRAYLIGHYT, and CLDICE that need to be applied to the OC data to ensure accuracy that can limit or completely exclude data, see Figure (3.3). Sentintel 3 data was already processed by NCCOS and the generated CIVs do not need the masks applied; however, the data still needed to be sifted through to locate and obtain viable days with viable data. The index value ranges from 0-254 with the number range from 0-250 indicating the intensity of the bloom, and the number range 251-254 indicate the presence of data eliminated by the masks, LAND, HIGLINT, STRAYLIGHT, and CLDICE. Once imported into SeaDAS a simple pixel inquiry was used to see if days were affected by these masks.

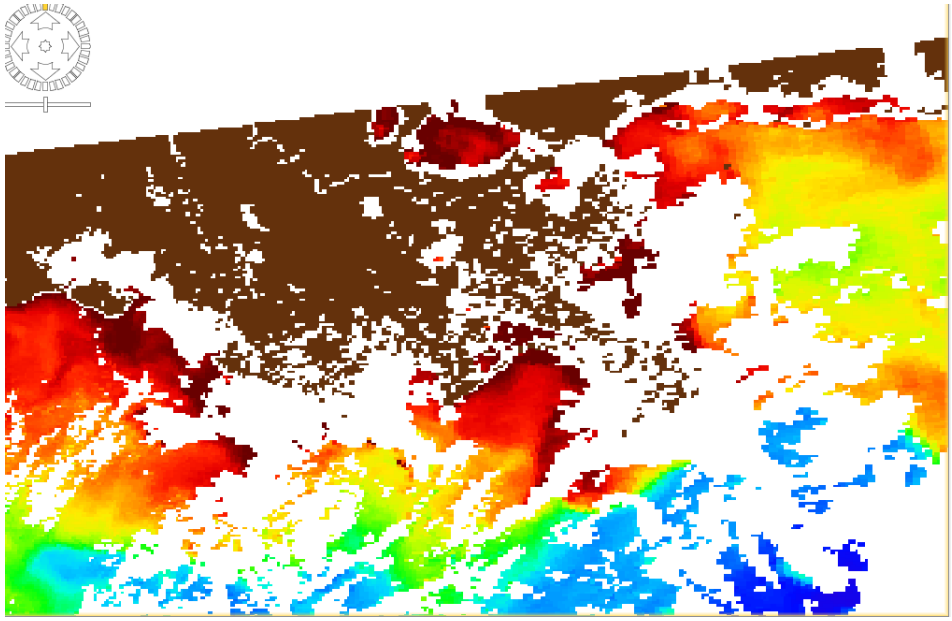

Figure 3.2. MODIS Chlor-a pixel data with only the Land mask applied

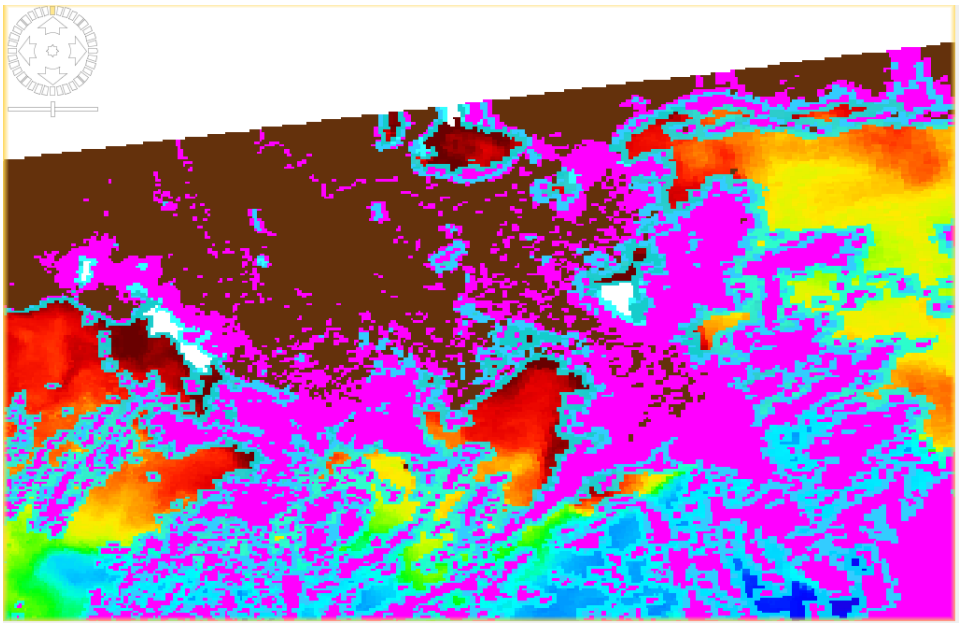

Figure 3.3. MODIS Chlor-a pixel data with Land, Higlint (gray), Straylight (turquoise blue), and CLDICE (fuchsia) masks applied

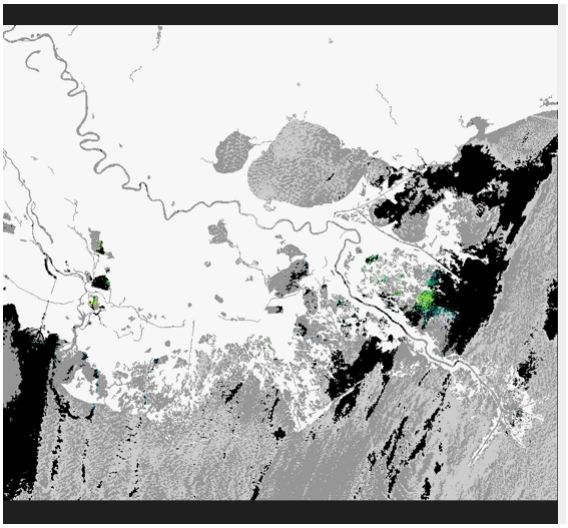

Figure 3.4. CIVs for a day with masks eliminating almost all the data in Lake Pontchartrain on January 2, 2021

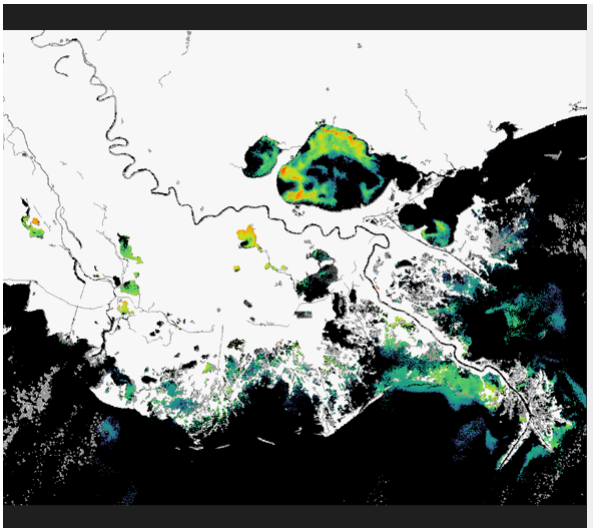

Figure 3.5. CIVs indicating a massive bloom in Lake Pontchartrain on July 27, 2021

Once the days with viable data have been identified the data must be extracted and saved as a csv file. Examples of the raw data are shown by Figures (3.6 & 3.7 & 3.8) To do this a mask shape file for Lake Pontchartrain was created in ArcGIS then imported into SeaDAS. For OC data, once in the program the other aforementioned masks are ap-

plied then manipulated with mask tools in SeaDAS, each pixel's data is extracted. For Sentinel 3 data the only additional mask needed was one that can extract the values 0-250.

The next step in data processing was to match the OC data with the corresponding day's SST data. When exporting the OC and SST data the latitude and longitude locations given are the left top corner and right bottom corner of the 1000x1000 m cell. The mid latitude and mid longitude were both calculated to give the location of the middle of the cell. However, this value was not used to match the cells are they are not a direct match as the swaths for OC and SST are shot at slightly different zenith angles. To match the OC data to the SST data, when imported into SeaDAS the program assigns each cell an x and y value, and since both OC and SST were reprojected into the same coordinate plane in SeaDAS these values are exact matches for both data sets. A Matlab code was written to find the matching OC and SST pixels and create a new csv file with the now matching appropriate data with all the needed MODIS-Aqua remote sensing variable data.

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pixel-X | Pixel-Y | Longitude | Latitude | aot_869 | angstrom | Rrs_412 | Rrs_443 | Rrs_469 | Rrs_488 | Rrs_531 | Rrs_547 | Rrs_555 | Rrs_645 | Rrs_667 | Rrs_678 | chlor_a | Kd_490 | pic | poc | ipar | nflh | par | l2_flags | longitude | latitude |
| 508.5 | 772.5 | -90.209 | 30.3444 | 0.1009 | 1.6284 | 0.002 | 0.00314 | 0.00339 | 0.0037 | 0.0054 | 0.00608 | 0.00593 | 0.00583 | 0.00536 | 0.00513 | 7.54925 | 0.5338 | NaN | 400 | 0.00203 | -0.0163 | 61.89 | 1.1E+09 | -90.203 | 30.344 |
| 509.5 | 772.5 | -90.195 | 30.3441 | 0.1009 | 1.6284 | 0.002 | 0.00314 | 0.00339 | 0.0037 | 0.0054 | 0.00608 | 0.00593 | 0.00583 | 0.00536 | 0.00513 | 7.54925 | 0.5338 | NaN | 400 | 0.00203 | -0.0163 | 61.89 | 1.1E+09 | -90.203 | 30.344 |
| 510.5 | 772.5 | -90.182 | 30.3438 | 0.1032 | 1.6728 | 0.00104 | 0.00215 | 0.00236 | 0.00266 | 0.00442 | 0.0052 | 0.00514 | 0.005 | 0.00474 | 0.0045 | 12.7098 | 1.0504 | NaN | 500 | 0.00202 | -0.1348 | 61.944 | 1.1E+09 | -90.182 | 30.3483 |
| 511.5 | 772.5 | -90.169 | 30.3435 | 0.0977 | 1.7843 | 3.04E-04 | 0.00145 | 0.00188 | 0.00225 | 0.00423 | 0.00515 | 0.00509 | 0.00473 | 0.00454 | 0.00422 | 20.1336 | 2.1848 | NaN | 746.6 | 0.00202 | -0.3099 | 61.674 | 1.1E+09 | -90.157 | 30.3402 |
| 512.5 | 772.5 | -90.156 | 30.3432 | 0.0977 | 1.7843 | 3.04E-04 | 0.00145 | 0.00188 | 0.00225 | 0.00423 | 0.00515 | 0.00509 | 0.00473 | 0.00454 | 0.00422 | 20.1336 | 2.1848 | NaN | 746.6 | 0.00202 | -0.3099 | 61.674 | 1.1E+09 | -90.157 | 30.3402 |
| 507.5 | 773.5 | -90.222 | 30.3333 | 0.0964 | 1.9724 | -0.0014 | -4.00E-06 | 4.22E-04 | 9.56E-04 | 0.00314 | 0.00396 | 0.0039 | 0.00398 | 0.00361 | 0.00343 | 74.9307 | NaN | NaN | NaN | 0.00201 | 0.01076 | 61.86 | 1.1E+09 | -90.221 | 30.3273 |
| 508.5 | 773.5 | -90.209 | 30.333 | 0.0916 | 1.7384 | 0.00209 | 0.00322 | 0.00348 | 0.00381 | 0.0057 | 0.00648 | 0.00628 | 0.00609 | 0.00558 | 0.00529 | 8.35526 | 0.6026 | NaN | 417.2 | 0.00203 | -0.0745 | 61.872 | 1.1E+09 | -90.2 | 30.3316 |
| 509.5 | 773.5 | -90.196 | 30.3327 | 0.0916 | 1.7384 | 0.00209 | 0.00322 | 0.00348 | 0.00381 | 0.0057 | 0.00648 | 0.00628 | 0.00609 | 0.00558 | 0.00529 | 8.35526 | 0.6026 | NaN | 417.2 | 0.00203 | -0.0745 | 61.872 | 1.1E+09 | -90.2 | 30.3316 |
| 510.5 | 773.5 | -90.183 | 30.3324 | 0.0936 | 1.8043 | 8.58E-04 | 0.00194 | 0.00234 | 0.00269 | 0.00463 | 0.0055 | 0.0054 | 0.00518 | 0.0049 | 0.00459 | 14.5338 | 1.2812 | NaN | 592.4 | 0.00202 | -0.2251 | 61.918 | 1.1E+09 | -90.179 | 30.3359 |
| 511.5 | 773.5 | -90.17 | 30.3321 | 0.0936 | 1.8043 | 8.58E-04 | 0.00194 | 0.00234 | 0.00269 | 0.00463 | 0.0055 | 0.0054 | 0.00518 | 0.0049 | 0.00459 | 14.5338 | 1.2812 | NaN | 592.4 | 0.00202 | -0.2251 | 61.918 | 1.1E+09 | -90.179 | 30.3359 |
| 512.5 | 773.5 | -90.156 | 30.3318 | 0.0908 | 1.8114 | 0.00115 | 0.00234 | 0.00267 | 0.00304 | 0.00503 | 0.00591 | 0.00576 | 0.00561 | 0.00531 | 0.00498 | 12.5138 | 1.0272 | NaN | 525.8 | 0.00203 | -0.1966 | 61.674 | 1.1E+09 | -90.154 | 30.3278 |
| 513.5 | 773.5 | -90.143 | 30.3316 | 0.1039 | 1.639 | 0.00124 | 0.00243 | 0.00282 | 0.00317 | 0.00534 | 0.00623 | 0.00605 | 0.00558 | 0.00527 | 0.00491 | 12.8629 | 1.0686 | NaN | 536.6 | 0.00203 | -0.2543 | 61.668 | 1.1E+09 | -90.133 | 30.3321 |
| 514.5 | 773.5 | -90.13 | 30.3313 | 0.1039 | 1.639 | 0.00124 | 0.00243 | 0.00282 | 0.00317 | 0.00534 | 0.00623 | 0.00605 | 0.00558 | 0.00527 | 0.00491 | 12.8629 | 1.0686 | NaN | 536.6 | 0.00203 | -0.2543 | 61.668 | 1.1E+09 | -90.133 | 30.3321 |
| 516.5 | 773.5 | -90.104 | 30.3307 | 0.1252 | 1.5888 | -2.62E-04 | 0.0012 | 0.00165 | 0.00224 | 0.00446 | 0.00534 | 0.00529 | 0.0051 | 0.0048 | 0.00453 | 22.4973 | 2.671 | NaN | 945.6 | 0.00201 | 0.0134 | 61.518 | 1.1E+09 | -90.109 | 30.324 |
| 507.5 | 774.5 | -90.222 | 30.3219 | 0.0964 | 1.9724 | -0.0014 | -4.00E-06 | 4.22E-04 | 9.56E-04 | 0.00314 | 0.00396 | 0.0039 | 0.00398 | 0.00361 | 0.00343 | 74.9307 | NaN | NaN | NaN | 0.00201 | 0.01076 | 61.86 | 1.1E+09 | -90.221 | 30.3273 |
| 508.5 | 774.5 | -90.209 | 30.3216 | 0.0893 | 2.0069 | -5.54E-04 | 8.56E-04 | 0.00125 | 0.00183 | 0.00403 | 0.00484 | 0.00475 | 0.00488 | 0.00447 | 0.00423 | 29.5459 | 4.6024 | NaN | NaN | 0.00202 | 0.01144 | 61.298 | 1.1E+09 | -90.218 | 30.3149 |

Figure 3.6. Raw MODIS-Aqua OC data after being imported into a csv file

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pixel-X | Pixel-Y | Longitude | Latitude | sst | l2_flags | sstref | qual_sst | flags_sst | bias_sst | stdv_sst | longitude | latitude |
| 509.5 | 775.5 | -90.1966 | 30.30992 | 31.64 | NaN | 29.125 | NaN | NaN | 0.115 | 0.525 | -90.193 | 30.30687 |
| 510.5 | 775.5 | -90.1835 | 30.30964 | 31.64 | NaN | 29.125 | NaN | NaN | 0.115 | 0.525 | -90.193 | 30.30687 |
| 511.5 | 775.5 | -90.1704 | 30.30935 | 31.82 | NaN | 29.125 | NaN | NaN | 0.115 | 0.525 | -90.1718 | 30.31118 |
| 513.5 | 775.5 | -90.1442 | 30.30878 | 31.945 | NaN | 29.13 | NaN | NaN | 0.115 | 0.525 | -90.1472 | 30.30313 |
| 509.5 | 776.5 | -90.1969 | 30.29854 | 32.1 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1896 | 30.29449 |
| 510.5 | 776.5 | -90.1838 | 30.29826 | 32.1 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1896 | 30.29449 |
| 511.5 | 776.5 | -90.1707 | 30.29798 | 31.895 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1684 | 30.29882 |
| 512.5 | 776.5 | -90.1576 | 30.2977 | 31.895 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1684 | 30.29882 |
| 513.5 | 776.5 | -90.1445 | 30.29741 | 31.945 | NaN | 29.13 | NaN | NaN | 0.115 | 0.525 | -90.1472 | 30.30313 |
| 514.5 | 776.5 | -90.1314 | 30.29712 | 31.89 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1228 | 30.29509 |
| 515.5 | 776.5 | -90.1183 | 30.29684 | 31.89 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1228 | 30.29509 |
| 508.5 | 777.5 | -90.2103 | 30.28745 | 32.22 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.2109 | 30.29014 |
| 509.5 | 777.5 | -90.1972 | 30.28717 | 32.1 | NaN | 29.115 | NaN | NaN | 0.115 | 0.525 | -90.1896 | 30.29449 |
| 510.5 | 777.5 | -90.1841 | 30.28689 | 32.46 | NaN | 29.11 | NaN | NaN | 0.115 | 0.525 | -90.1862 | 30.2821 |

Figure 3.7. Raw MODIS-Aqua SST data after being imported into a csv file

| A | B | C | D | E |
|---|---|---|---|---|
| Pixel-X | Pixel-Y | Longitude | Latitude | band_1 |
| 1695.5 | 145.5 | -90.2077 | 30.38308 | 0 |
| 1696.5 | 145.5 | -90.2045 | 30.38301 | 0 |
| 1697.5 | 145.5 | -90.2014 | 30.38294 | 0 |
| 1698.5 | 145.5 | -90.1983 | 30.38287 | 0 |
| 1699.5 | 145.5 | -90.1952 | 30.38281 | 0 |
| 1700.5 | 145.5 | -90.1921 | 30.38274 | 0 |
| 1701.5 | 145.5 | -90.189 | 30.38267 | 0 |
| 1694.5 | 146.5 | -90.2109 | 30.38044 | 0 |
| 1695.5 | 146.5 | -90.2077 | 30.38037 | 0 |
| 1696.5 | 146.5 | -90.2046 | 30.3803 | 0 |
| 1697.5 | 146.5 | -90.2015 | 30.38024 | 0 |
| 1698.5 | 146.5 | -90.1984 | 30.38017 | 0 |

Figure 3.8. Raw CIV data after being imported into a csv file

Now that the OC and SST data are together matching the CIV to the appropriate cell is needed so that the model can be trained to predict the index value. The first major issue that was encountered was the CIV data pixels are 300x300 m and MODIS-Aqua data pixels are 1000x1000 m meaning that it could not be as simple as matching the cells as it was with the OC and SST data. To do this the data was resampled with a Matlab code and an average index value was assigned for the CIV pixels that fall within each of the MODIS-Aqua pixels. In order for Matlab to process the data mathematically the mid latitude and longitude calculated previously had to be converted from decimal degrees to meters for every data set then converted back again for the final data set. As previously mentioned, a large driving point of this thesis is to consider lag time between MODIS-Aqua data and CIV data. This day lag data set is created by running the above-

mentioned Matlab code for the correct lag day, for example a 15-day lag MODIS-Aqua data on 7/1/21 would be matched with CIV data on 7/16/21. All the data that falls into a specific day lag were combined at the end to create the larger complete day lag set. For example, the 15-day lag data set had a possible 71 smaller data sets from 2021 to 2022. It is only possible because not all viable MODIS-Aqua data has a viable CIV data counterpart for that specific day lag. See Figure (3.11), the red blocks indicate days that are not viable and green means there is possible data matching. In addition, just because there is matching lag day data available does not mean that the two data sets actually have a significant amount of data that coincide with one another.

| | Modis | 1 day | 2 day | 3 day | 4 day | 5 day | 6 day | 7 day |
|---|---|---|---|---|---|---|---|---|
| 1 | 1_2 | 1/3/2021 | 1/4/2021 | 1/5/2021 | 1/6/2021 | 1/7/2021 | 1/8/2021 | 1/9/2021 |
| 2 | 1_3 | 1/4/2021 | 1/5/2021 | 1/6/2021 | 1/7/2021 | 1/8/2021 | 1/9/2021 | 1/10/2021 |
| 3 | 1_4 | 1/5/2021 | 1/6/2021 | 1/7/2021 | 1/8/2021 | 1/9/2021 | 1/10/2021 | 1/11/2021 |
| 4 | 1_5 | 1/6/2021 | 1/7/2021 | 1/8/2021 | 1/9/2021 | 1/10/2021 | 1/11/2021 | 1/12/2021 |
| 5 | 1_9 | 1/10/2021 | 1/11/2021 | 1/12/2021 | 1/13/2021 | 1/14/2021 | 1/15/2021 | 1/16/2021 |
| 6 | 1_13 | 1/14/2021 | 1/15/2021 | 1/16/2021 | 1/17/2021 | 1/18/2021 | 1/19/2021 | 1/20/2021 |
| 7 | 1_15 | 1/16/2021 | 1/17/2021 | 1/18/2021 | 1/19/2021 | 1/20/2021 | 1/21/2021 | 1/22/2021 |
| 8 | 1_16 | 1/17/2021 | 1/18/2021 | 1/19/2021 | 1/20/2021 | 1/21/2021 | 1/22/2021 | 1/23/2021 |
| 9 | 4_1 | 4/2/2021 | 4/3/2021 | 4/4/2021 | 4/5/2021 | 4/6/2021 | 4/7/2021 | 4/8/2021 |
| 10 | 6_13 | 6/14/2021 | 6/15/2021 | 6/16/2021 | 6/17/2021 | 6/18/2021 | 6/19/2021 | 6/20/2021 |
| 11 | 7_21 | 7/22/2021 | 7/23/2021 | 7/24/2021 | 7/25/2021 | 7/26/2021 | 7/27/2021 | 7/28/2021 |
| 12 | 7_22 | 7/23/2021 | 7/24/2021 | 7/25/2021 | 7/26/2021 | 7/27/2021 | 7/28/2021 | 7/29/2021 |
| 13 | 7_23 | 7/24/2021 | 7/25/2021 | 7/26/2021 | 7/27/2021 | 7/28/2021 | 7/29/2021 | 7/30/2021 |
| 14 | 7_24 | 7/25/2021 | 7/26/2021 | 7/27/2021 | 7/28/2021 | 7/29/2021 | 7/30/2021 | 7/31/2021 |
| 15 | 7_27 | 7/28/2021 | 7/29/2021 | 7/30/2021 | 7/31/2021 | 8/1/2021 | 8/2/2021 | 8/3/2021 |
| 16 | 7_29 | 7/30/2021 | 7/31/2021 | 8/1/2021 | 8/2/2021 | 8/3/2021 | 8/4/2021 | 8/5/2021 |
| 17 | 8_7 | 8/8/2021 | 8/9/2021 | 8/10/2021 | 8/11/2021 | 8/12/2021 | 8/13/2021 | 8/14/2021 |
| 18 | 8_14 | 8/15/2021 | 8/16/2021 | 8/17/2021 | 8/18/2021 | 8/19/2021 | 8/20/2021 | 8/21/2021 |
| 19 | 8_16 | 8/17/2021 | 8/18/2021 | 8/19/2021 | 8/20/2021 | 8/21/2021 | 8/22/2021 | 8/23/2021 |

Figure 3.9. An example of one of the sections of the organization chart showing the matching CIV days with the corresponding MODIS-Aqua OC

### 3.3. Development of The HAB Forecasting Models

Before any analysis of each day lag model can be assessed, the algorithm needs to be chosen and its base parameters changed to improve model performance. The data sets were imported into an open-source machine learning platform called WEKA and XGBoost

was used in R Studio using an R script. In WEKA various model building machine learning algorithms can be employed and their parameters easily modified without having to deal with the coding aspect to determine which algorithm is most suitable. From multiple experimentation with several algorithms available in WEKA such as multiple perceptron, gaussian processes, additive regression (a form of AdaBoost), and random forest, the RF model performed the best. For each model WEKA will generate a performance metric containing the parameters and statistical measurements used to evaluate the model's performance. These parameters and measurements are the correlation coefficient ($R^2$), the mean absolute error (MAE) and the root mean square error (RMSE). The closer the $R^2$ value is to one the greater the model's performance , and when MAE and RMSE are all close to zero the greater the model's performance. Using these metrics, the RF and XGBoost models were chosen to move forward with as they performed the best. XGBoost was used in R Studio as WEKA does not have that algorithm in its database. When comparing the XGBoost statistics to RF statistics they were found to be comparable and either algorithm could have been used to move forward. Random Forest was ultimately chosen as using it within the WEKA platform proved to be more user friendly, easier to modify parameters, and easier to save and load each day lag model.

Optimizing the RF model in WEKA was the next step in the process. When testing any of the algorithms in WEKA there are three different testing methods available. First there is the training/test set method which is the most basic, where the data is split into a training and test data by the user prior to importing into WEKA. The training model is not immediately evaluated within WEKA, instead the evaluation metrics are assessed using the supplied test set in a separate step. This method is rudimentary and not

very robust. The next option is the percentage split testing where a set percentage of a data set is split for training and the rest is used for testing. Typically a split of 60% training and 40% testing is used when assessing the model and an evaluation metrics are given immediately. The percentage split is essentially the same as the first method described but there is less user control over the testing and training data set. The third option is the cross-validation method which is the most robust method. For this method the data is split into sections called folds and the model is trained with the majority of the data and tested for each fold. It is easier to imagine it as a circle or pie where the number of folds is how many slices of pie there are. A slice of pie is taken out for testing and a model is developed using the remaining pieces. After testing with the first slice of pie, it is put back into the whole and a different slice is taken out and the process is repeated using the new slice as testing and the remaining data as training. The model is then improved each time for each fold, so a data set with 10 folds would be trained and improved 10 times. This technique is especially effective with large data sets as it allows the algorithm to be tested against the entire data set at some point in developing the algorithm process. Due to its robust approach the cross-validation method was chosen. In addition to the cross-validation method a third of the entire data set for each lag day will be put aside for validation later to test the model on data it hasn't seen before.

Now that the base training and test method has been chosen, tweaking the base parameters for the RF model was needed. For the readers reference a user guide to the WEKA platform is included in Appendix B. There are many different parameters available for adjustment for the Random Forest model. After systematically changing all the properties to varying numbers the only significantly impactful number that improved the

model's performance metric when changed from the default setting was the number of iterations parameter. The iterations parameter is the number of trees generated in the forest, denoted by an I and numIterations section. Other parameters such as the M and V parameter which respectively stand for the minimum number of instances per leaf and minimum numeric class variance proportion of the training variance for the split. These parameters had a slight impact but not enough impact to be considered significant. With all that taken into consideration, all parameters except the number of iterations and number of folds were kept on the default setting. The Figures (3.13 & 3.14 & 3.15) below show the correlation coefficient for 1, 15, and 30 day lags at 50, 100, 150, 200, 250, 300, 400, and 500. The trend for each of the three graphs shows as the number of instances increases the correlation coefficient increases; as well as simultaneously showing as the number of folds increases the correlation coefficient increases. This trend is the strongest with the 1 day data set, and the weakest with the 15 day data set. However, the overall correlation coefficient is highest for the 15 day data set with the 50 fold 250 iteration being the highest of all the data points. The best performing combination was different for all three day lag data sets; however, since the 50 fold 250 iteration performed the best it was decided that it would be the optimal settings for the model.

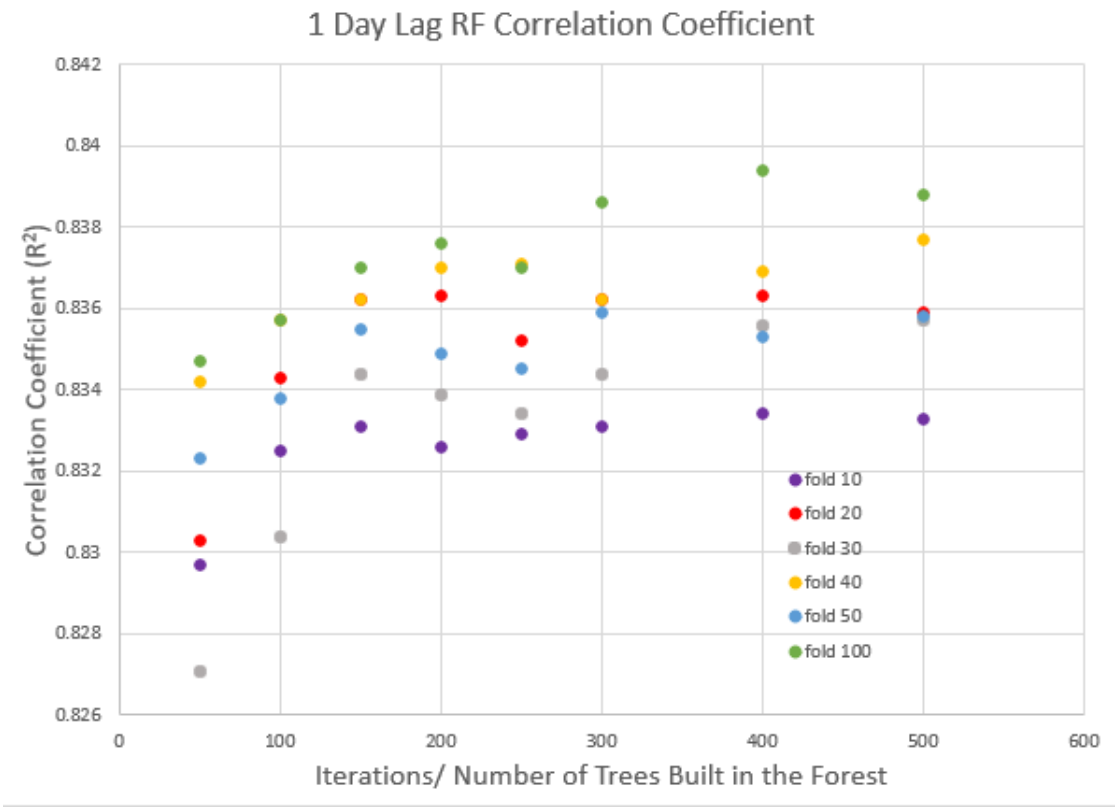


Figure 3.10. 1 day lag data set correlation coefficient of different model training parameters

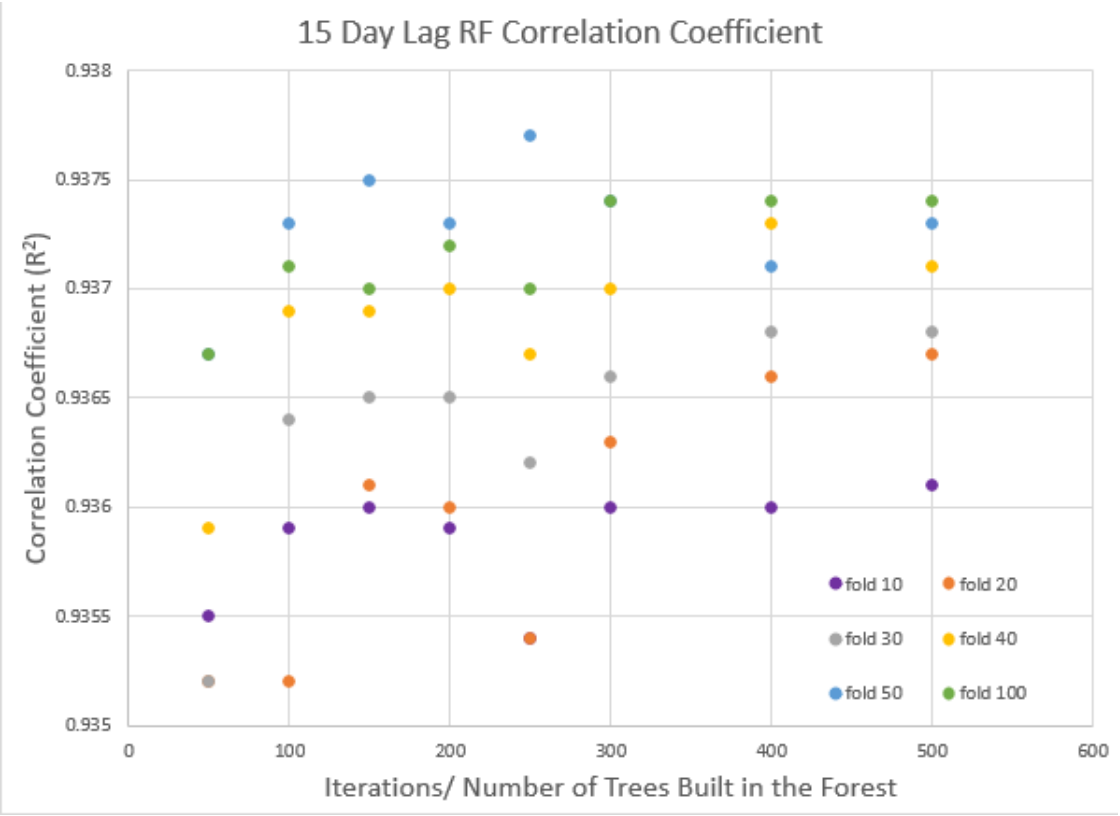


Figure 3.11. 15 day lag data set correlation coefficient of different model training parameters

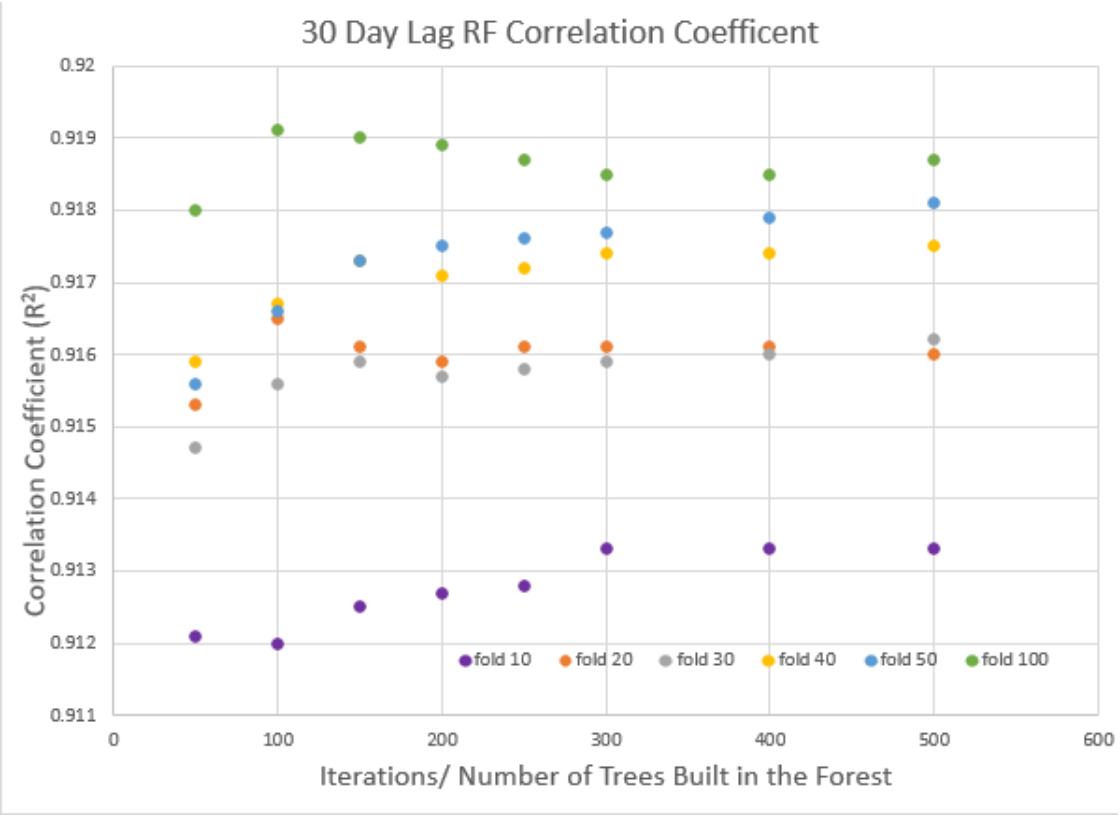


Figure 3.12. 30 lag day data set correlation coefficient of different model training parameters

The next step was to determine which day lags were the most effective in predicting the index value, day lags from 1 day to 30 days were created for the 2021 data set as that is the year of the most intense and frequent algal blooms. The day lag model's correlation coefficient was graphed from 1 to 30 days and it was determined the best day span was from the 15 day lag to the 22 day lag data sets. This can be shown by Figure (3.16) below, with a peak being shown at the 19 day lag and 20 day lag mark with the 6th day lag model being an outlier. Using the day lags 15 through 22, an eight day forecasting period using 8 different models was developed to predict CyanoHABs 15 to 22 days in advance.

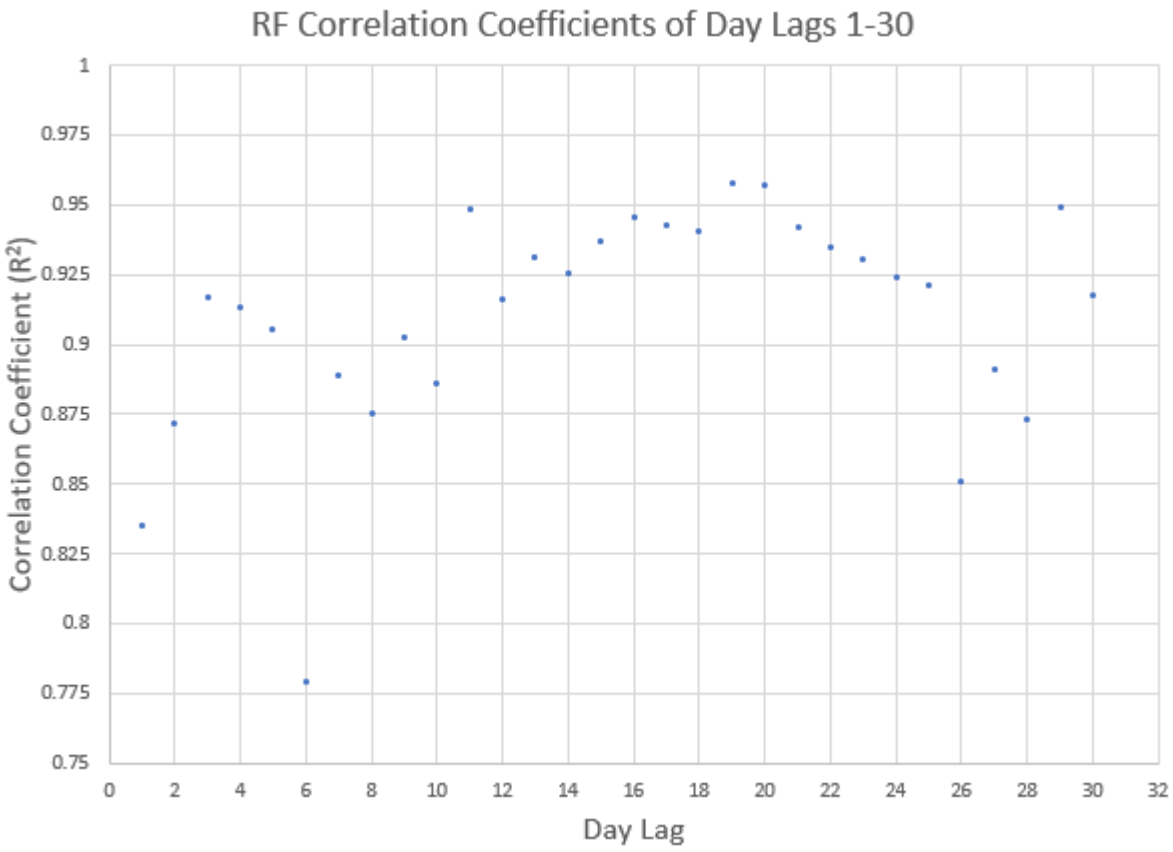


Figure 3.13. A chart showing the correlation coefficients of the day lag data sets from 1 to 30

With the optimal day lag range found, unnecessary data from 1-14 & 23-30 day lag can be excluded from the 2022 data set which is predominately the zero index value data set. The 2021 and the 2022 data sets are combined and were ready to be used in the algorithm. As stated previously, the data was split so that a third of the data was excluded for validation and mapping. This was done with another simple Matlab code. It is important to note that the data was organized chronologically with January 2021 data being first and December 2022 data being last. This does not affect the algorithm as the RF model randomly grabs from the training and testing data set when developing the model; how-

ever, this does allow for a chronological spread of data when splitting off into training and validation data sets allowing for seasonal analysis of the model. In addition, the mat lab code after splitting the data also removes the N/A data points as well as the algorithm does not see these as valid data.

The next step was to create the time series performance graphs and the observed versus predicted CIV maps to analyze the data chronologically and spatially. To do this a number range had to be assigned for each CIV day within the larger day lag data set. A chart was made, Figure (3.17), to help keep the data organized as well as easy to manage. Also, at this point it was realized that the latitude and longitude for each cell were omitted as they are not a variable that needs to be considered in the training of the model and needed to be added back to the data so it would be possible to map the predicted values later. This was done by simply tweaking the Matlab code that matches the CIV with the MODIS-Aqua OC data, and the day lag and validation split data sets were processed again. Since the order and magnitude of the data sets did not change, the latitude, longitude, and number assignment that dictated what index day the data points belonged to were simply added to the already predicted validation files. Since the validation data sets range in the 1,000s the average value for the observed and predicted value was determined for each index day. Then the average values were plotted against the increasing day to show how much they varied from each other as time progresses.

| | 19 day | Range | | | 20 day | Range | |
|---|---|---|---|---|---|---|---|
| | | beginning | end | | | beginning | end |
| 1/21/2021 | 1 | 1 | 400 | 1/25/2021 | 4 | 1 | 163 |
| 1/28/2021 | 5 | 401 | 758 | 1/29/2021 | 5 | 164 | 521 |
| 2/1/2021 | 6 | 759 | 1517 | 2/2/2021 | 6 | 522 | 1280 |
| 2/3/2021 | 7 | 1518 | 2210 | 2/4/2021 | 7 | 1281 | 1973 |
| 2/4/2021 | 8 | 2211 | 2628 | 7/3/2021 | 10 | 1974 | 1982 |
| 4/20/2021 | 9 | 2629 | 2742 | 8/10/2021 | 11 | 1983 | 2314 |
| 8/9/2021 | 11 | 2743 | 3074 | 8/11/2021 | 12 | 2315 | 2694 |
| 8/10/2021 | 12 | 3075 | 3454 | 8/12/2021 | 13 | 2695 | 2873 |
| 8/11/2021 | 13 | 3455 | 3633 | 8/13/2021 | 14 | 2874 | 2916 |
| 8/12/2021 | 14 | 3634 | 3675 | 8/16/2021 | 15 | 2917 | 3141 |
| 8/15/2021 | 15 | 3676 | 3901 | 8/18/2021 | 16 | 3142 | 3154 |
| 8/17/2021 | 16 | 3902 | 3914 | 9/3/2021 | 18 | 3155 | 3694 |
| 8/26/2021 | 17 | 3915 | 4241 | 9/5/2021 | 19 | 3695 | 3944 |
| 9/2/2021 | 18 | 4242 | 4781 | 9/9/2021 | 20 | 3945 | 4206 |
| 9/4/2021 | 19 | 4782 | 5031 | 9/10/2021 | 21 | 4207 | 4388 |
| 9/9/2021 | 21 | 5032 | 5213 | 9/11/2021 | 22 | 4389 | 4464 |
| 9/10/2021 | 22 | 5214 | 5289 | 9/12/2021 | 23 | 4465 | 4880 |
| 9/11/2021 | 23 | 5290 | 5705 | 12/5/2021 | 25 | 4881 | 5558 |
| 9/13/2021 | 24 | 5706 | 6069 | 12/8/2021 | 27 | 5559 | 5910 |
| 12/4/2021 | 25 | 6070 | 6747 | 12/13/2021 | 28 | 5911 | 6108 |
| 12/5/2021 | 26 | 6748 | 7161 | 12/30/2021 | 32 | 6109 | 6261 |
| 12/12/2021 | 28 | 7162 | 7359 | 1/27/2022 | 1 | 6262 | 6453 |

Figure 3.14. An example of a portion of the organization chart that shows the dates that correlate to the appropriate number section

The spatial maps were created with the validation data set, and days that show visable blooms and have the most data points were chosen to be displayed. Two maps were created to compare, one that shows the observed index value and another that is its spatial clone that shows the model's predicted index value. The maps were created in ArcMAP using the Kriging tool to interpolate the points given. The maps were created using the same data sheets as the temporal data, but use all the points instead of the average. Unfortunately, the available data does not cover the entire lake, and the validation data only includes one third of the overall available data so the the interpolation values are limited. However, the purpose of these maps is to show the difference in observed data and the predicted values not to give high resolution images.

# Chapter 4. Results and Discussion

The following sections will go over the statistical, temporal, and spatial analysis of the eight different models developed to forecast harmful algal blooms. The models output is a predicted index value, not a concentration value. This index value was created by NCCOS to monitor if an algal bloom in Lake Pontchartrain is toxic or not i.e the presence of Cyanobacteria. Not all algal blooms in Lake Pontchartrain are toxic, and may not be detected by the monitoring system or may be given a low index value, but overall the monitoring system has historically performed well. However, the monitoring system is just that, as it only shows values for days that have already occurred or at the earliest a day after. While useful to know if a current bloom in the lake is toxic, it is not useful to preemptively warn the public or enough to implement counter measures to mitigate the bloom. The results of this thesis show that using the MODIS-Aqua satellite remote sensing data that a forecast 15 to 22 days in advance can be achieved. This is significant as currently an ability to forecast an index value in Lake Pontchartrain does not exist on the NCCOS website. Hopefully with this new contribution to science a dependable model can be implemented to give the needed tools to improve the overall health of Lake Pontchartrain.

## 4.1. Model Performance Metrics

Below are two tables. Table 4.1 shows the training statistics and Table 4.2 shows the validation statistics of the eight models labeled: 15 day, 16 day, 17 day, 18 day, 19 day, 20 day, 21 day, and 22 day. Each model is labeled after how many days in advance that the model can predict an index value. Three statistical metrics are shown: $R^2$, MAE, and

RMSE as well as the number of instances, which is how many sets of data were used for the training. Each metric carries a separate meaning but MAE and RMSE measure model error while $R^2$ measures the relationship between variables. For the RF models generated for this thesis the $R^2$ statistic measures the strength and direction of the models predicted values versus the models observed values. The closer a $R^2$ value is to 1 for a model the better the model has performed. It means that for the data set as a whole when the observed value is relatively high the predicted value is relatively high, and when the observed value is relatively low the predicted value is relatively low. It is a good measure that the model is able to predict high as well as low values. As it can be seen in Table 4.1 all the models produced a $R^2$ value over .91 meaning the model overall predicts low as well as high values well. The highest $R^2$ training value was generated by the 20 day model at .9341 and the lowest value was generated by the 22 day model at .9106.

It is important to note that while the $R^2$ is indicative of a good performing model it is not a good metric on evaluating the model's accuracy. The metrics MAE and RMSE are a better evaluation of model accuracy. As mentioned above these metrics evaluate the error of the model, but each in a different way. The MAE measures the mean absolute error of the model, meaning it shows the mean absolute numerical difference between the predicted and observed values of the model. For example from Table 4.1 the 20 day model had a 8.0372 MAE training value; meaning that for the entire data set on average the model's predicted value was plus or minus 8.0372 of the observed value. Naturally a lower MAE means that the model predicts closer to the observed value. When assessing this metric the magnitude of the observed value must be taken into account. For example the index value being predicted by these models ranges from 0 to 250. An observed value

of 100 and a predicted value of 108 or 92 puts the predicted value well within a 3.2% difference range of the observed value. While if the index value ranged from 0-50 the same 8 point difference falls within a 16 % difference range of the observed data, which is a significant change. Overall considering the 0-250 index value range a MAE value below 10 could be indicative of a well performing model, and any MAE value below 5 would be indicative of an exceptional model. As it can be seen by Table 4.1 all models produced an MAE below 10 with the 18 Day model being the highest at 9.1882 and the 19 Day model being the lowest at 6.2805.

RMSE is similar to MAE but it is the root mean squared error, where the error is squared. This is a good metric to consider as squaring the error gives a higher weight to large errors. For instance an error of 20 would turn into an error of 400 under the RMSE metric. A model with a relatively low MAE but a high RMSE value would indicate a large error outlier. For this reason a model with a RMSE that is more than triple the MAE would indicate a poor performing model. From Table 4.1 it can be seen that all the models had a training RMSE below 20 and all were generally double the MAE indicating a medium level variance in errors. Overall the training RMSE of the models was acceptable with the lowest being 13.7244 for the 16 Day model and the highest being 18.2096 for the 18 Day model.

The last metric is the instance value which is the number of rows or sets in the training data sets. This isn't as much of an important metric as the previously three described. However, it does give a bit of insight on the sample size. A study done on a small sample size carries less weight than one with a large sample size. The same logic applies to instances, when training models with more instances the more data the model sees giving

more validity to the three previous metrics.

When taking all four metrics into account, for training all models performed at an acceptable if not exceptional level. The best performing model using the training data set would be the 16 Day model as it has the most instances, the best RMSE, second best MAE, and third best $R^2$ values.

Table 4.1. Model Training Performance Metrics

| | 15 Day | 16 Day | 17 Day | 18 Day | 19 Day | 20 Day | 21 Day | 22 Day |
|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.9269 | 0.9289 | 0.9116 | 0.9239 | 0.9321 | 0.9341 | 0.9239 | 0.9106 |
| MAE | 7.0441 | 6.2989 | 7.8539 | 9.1882 | 6.2805 | 8.0372 | 7.8672 | 7.8985 |
| RMSE | 14.3857 | 13.7244 | 15.6485 | 18.2096 | 14.3862 | 15.9238 | 16.5932 | 15.9902 |
| Instances | 8972 | 9496 | 9622 | 7386 | 7912 | 7206 | 5742 | 6828 |

Table 4.2. Validation of Model Performance Metrics

| Metric | 15 Day | 16 Day | 17 Day | 18 Day | 19 Day | 20 Day | 21 Day | 22 Day |
|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.928 | 0.9336 | 0.9235 | 0.9345 | 0.9373 | 0.8227 | 0.9315 | 0.9171 |
| MAE | 6.7767 | 5.9101 | 7.174 | 8.3381 | 5.8738 | 12.5987 | 7.2988 | 7.4426 |
| RMSE | 14.226 | 13.1109 | 14.4477 | 16.775 | 13.9338 | 25.3981 | 15.8222 | 15.3411 |
| Instances | 4486 | 4748 | 4810 | 3692 | 3956 | 3602 | 2871 | 3413 |

While training model metrics are an important aspect of evaluation, validation of the model is even more important. To validate the model a third of the data was put aside to test the model on data that it has never seen before. Table 4.2 shows the the validation metrics to be compared to the training metrics. Ideally for a well performing model,

the validation metrics should match or out-perform the training model's metrics. It is very unlikely for the validation metrics to match the training exactly; therefore, a change in $R^2$ values less than .01, and a change in MAE and RMSE values less than 1 will be seen as not significant. The change in instance values are not taken into account as by definition the validation data set is a third of the entire data . With that in mind it can be seen that the 15 Day, 16 Day, 17 Day 18 Day, 19 Day, 21 Day, and 22 Day models performed the same as or better in all metrics than they did during training. The 20 Day model is the only model that had worse validation metrics in relation to its training metrics. Even with this down grade in statistical metrics of the 20 day model, overall it can be deduced by analyzing and comparing the two Tables 4.1 & 4.2 that statistically the performance is acceptable for making a forecast of CyanoHABs in Lake Pontchartrain 15 - 22 days in advance.

### 4.2. Model Forecast Time Series Graphs

The following figures below show a time series comparison, from January 2021 to December 2022, with the observed values depicted by the blue dots and the predicted values depicted by the orange line. Each day lag data set was split geographically into North and South sections of the lake not only to decrease cluttering of the graph, but to give a slight insight into spatial analysis of the lake. These graphs show how well the models predict at different times of the year, and can give insight on what times of the year it performs well or poorly.

The first graph, Figure 4.1, depicts the 15 day lag model. Both the north and south areas of the lake predict the major bloom in the Summer of 2021 and the smaller

bloom in the fall of 2022. Overall the model under predicts the index values for the summer bloom. For the fall bloom the model's predicted values match the observed values more closely. Spatially the blooms show that they have an approximate equal magnitude of occurrence in the north and south of the lake.

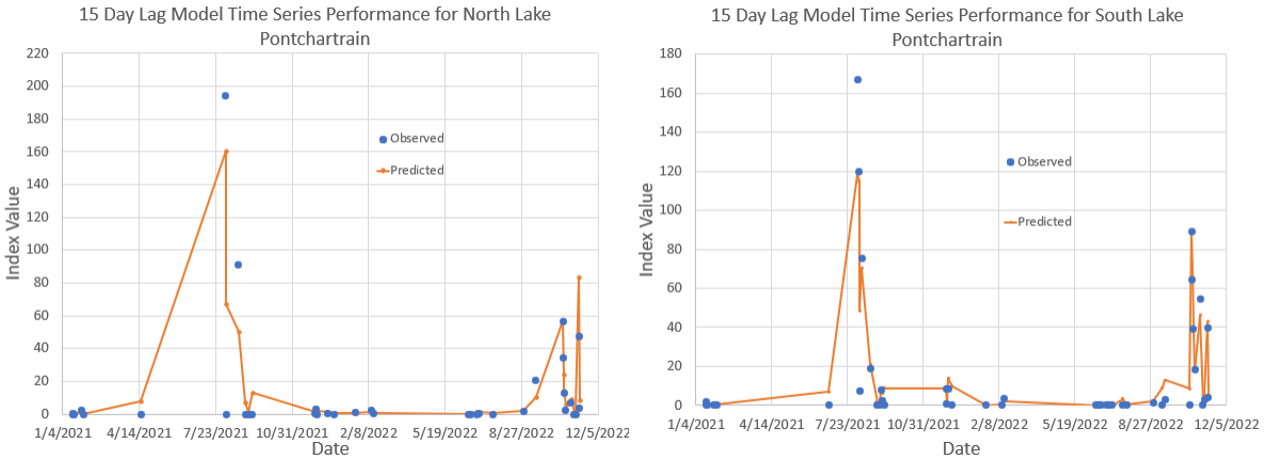


Figure 4.1. Two time series graphs for the 15 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.2 displays two graphs for the 16 day lag model one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the Summer of 2021 and the smaller bloom in the fall of 2022. The model closely predicts the values for both blooms, especially for the larger summer bloom in the northern part of the lake. Spatially the blooms show an approximately equal magnitude for the summer bloom in both the north and south part of the lake. The fall bloom shows a greater magnitude in the southern part of the lake.

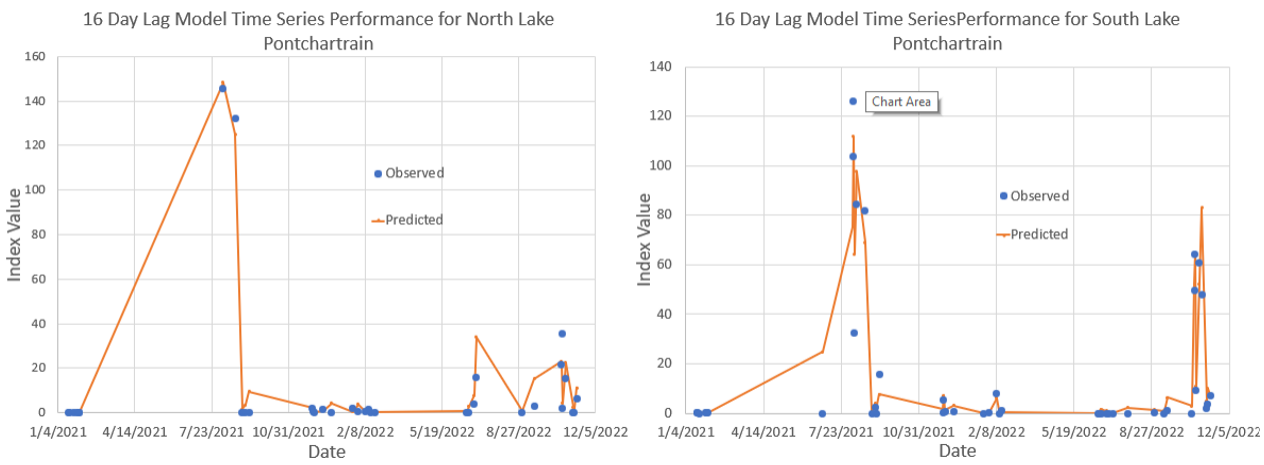


Figure 4.2. Two time series graphs for the 16 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.3 displays two graphs for the 17 day lag model one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the summer of 2021 with minor error. Spatially the blooms show an equal magnitude for the summer bloom in both the north and south part of the lake. The fall bloom shows a greater magnitude in the southern part of the lake.

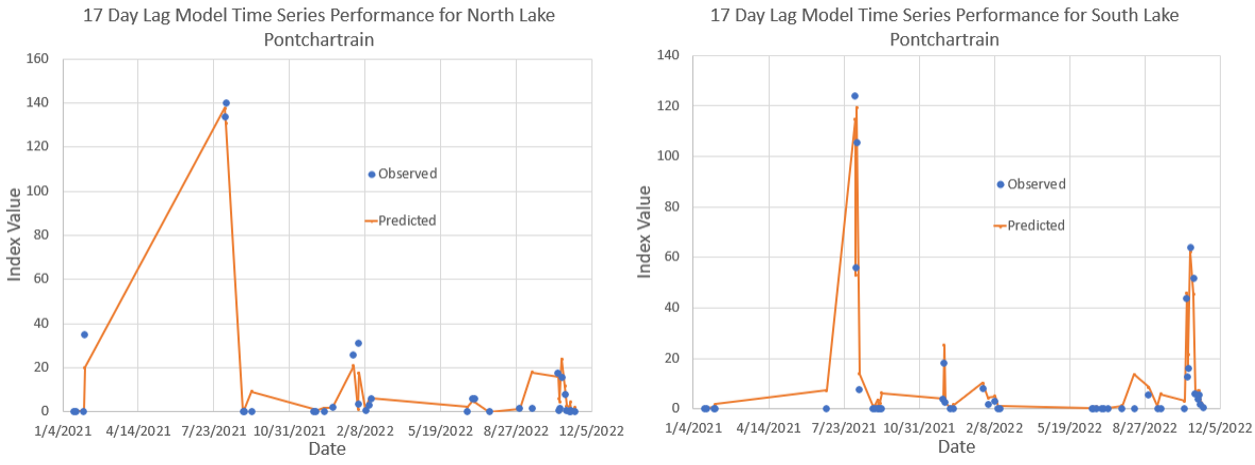


Figure 4.3. Two time series graphs for the 17 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.4 displays two graphs for the 18 day lag model, one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the summer of 2021 and the smaller bloom in the fall with a slight under prediction for both. Spatially the blooms show an equal magnitude for the summer bloom in both the north and south part of the lake. The fall bloom shows a significantly greater magnitude in the southern part of the lake.

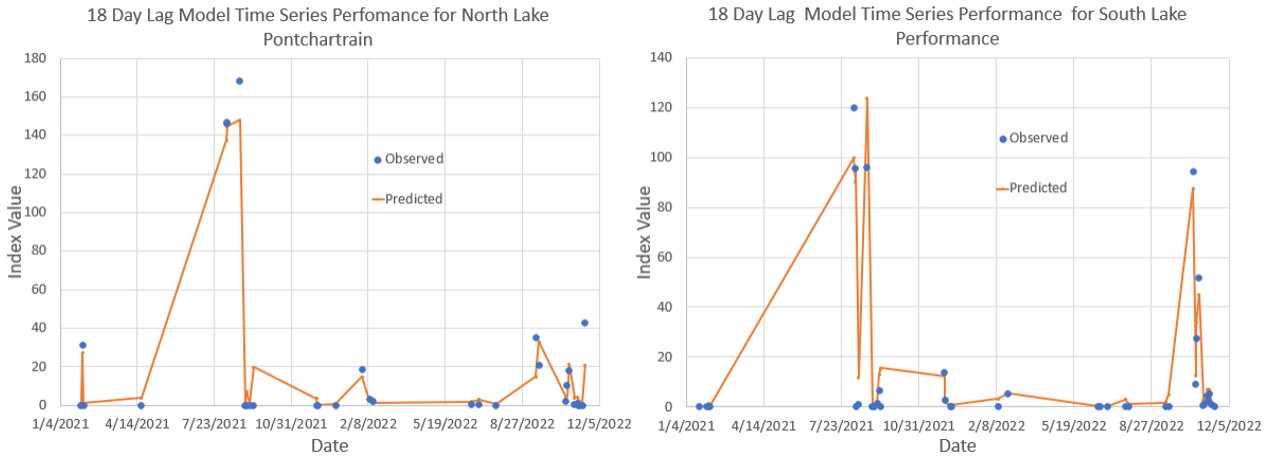


Figure 4.4. Two time series graphs for the 18 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.5 displays two graphs for the 19 day lag model, one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the summer of 2021 and the smaller bloom in the fall with a slight under prediction for both blooms. Spatially the blooms show a greater magnitude bloom in the north versus the south for the summer bloom. It may not appear so, but both charts show a near equivalent magnitude bloom for the fall bloom.

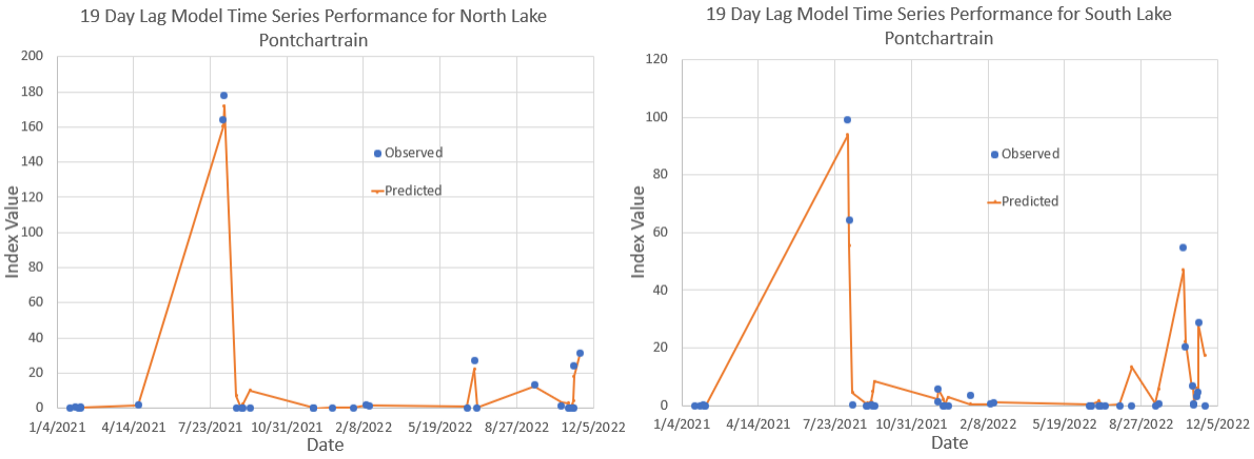


Figure 4.5. Two time series graphs for the 19 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.6 displays two graphs for the 20 day lag model, one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the summer of 2021 with a slight under prediction for the south portion of the lake. Spatially the blooms show a greater magnitude bloom in the north versus the south for the summer bloom. The fall bloom shows a greater magnitude in the southern part of the lake with a slight under prediction.

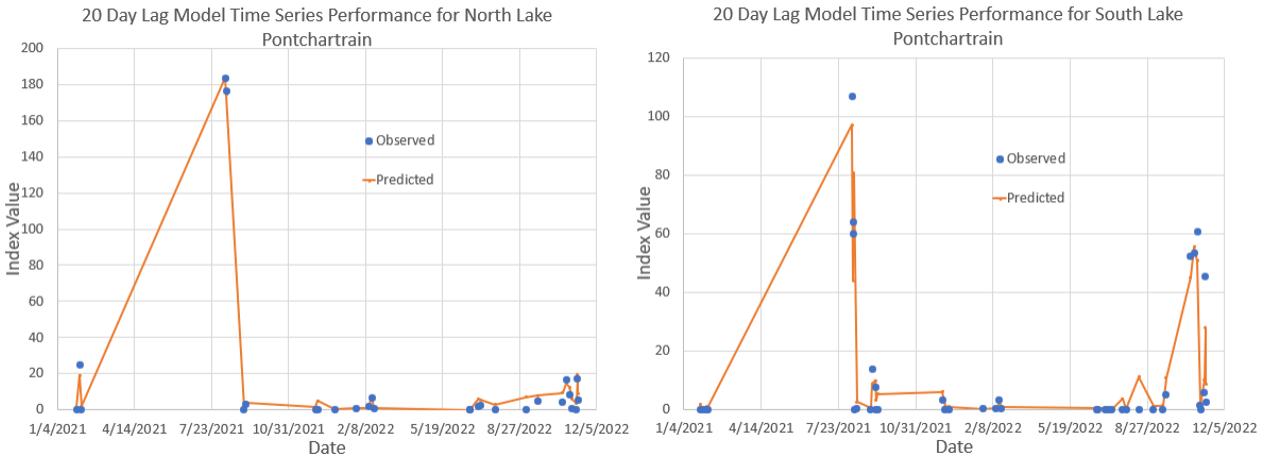


Figure 4.6. Two time series graphs for the 20 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.7 displays two graphs for the 21 day lag model, one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the summer of 2021 nearly perfect for both. Spatially the blooms show a significantly greater magnitude bloom in the north versus the south for the summer bloom. The fall bloom shows a greater magnitude in the southern part of the lake with a slight under prediction.

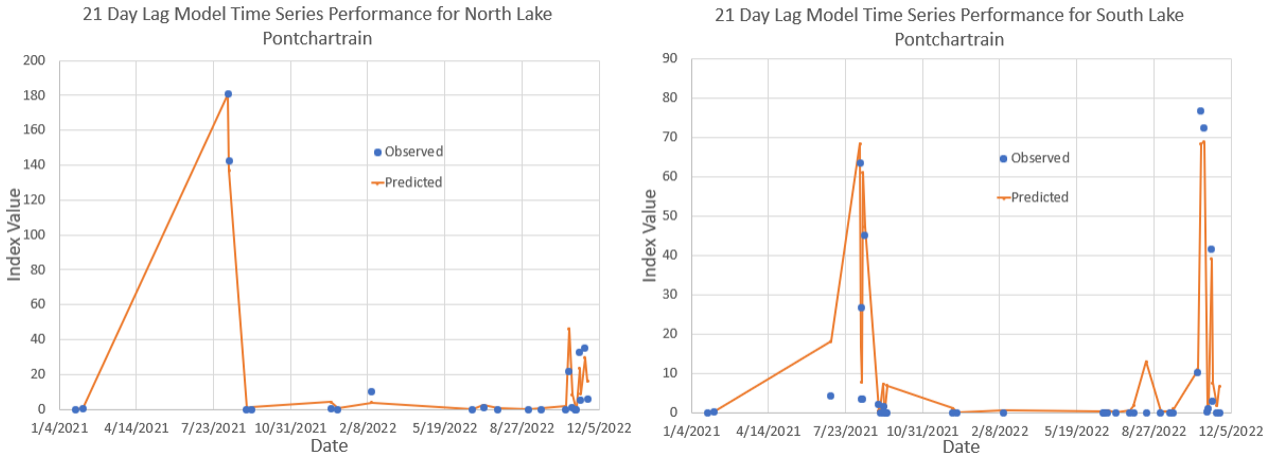


Figure 4.7. Two time series graphs for the 21 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Figure 4.8 displays two graphs for the 22 day lag model, one for the north and one for the south. Both the north and south areas of the lake predict the major bloom in the summer of 2021 nearly perfect for the north. Spatially the blooms show a significantly greater magnitude bloom in the north versus the south for the summer bloom. The fall bloom shows a greater magnitude in the southern part of the lake with a slight under prediction in the north.

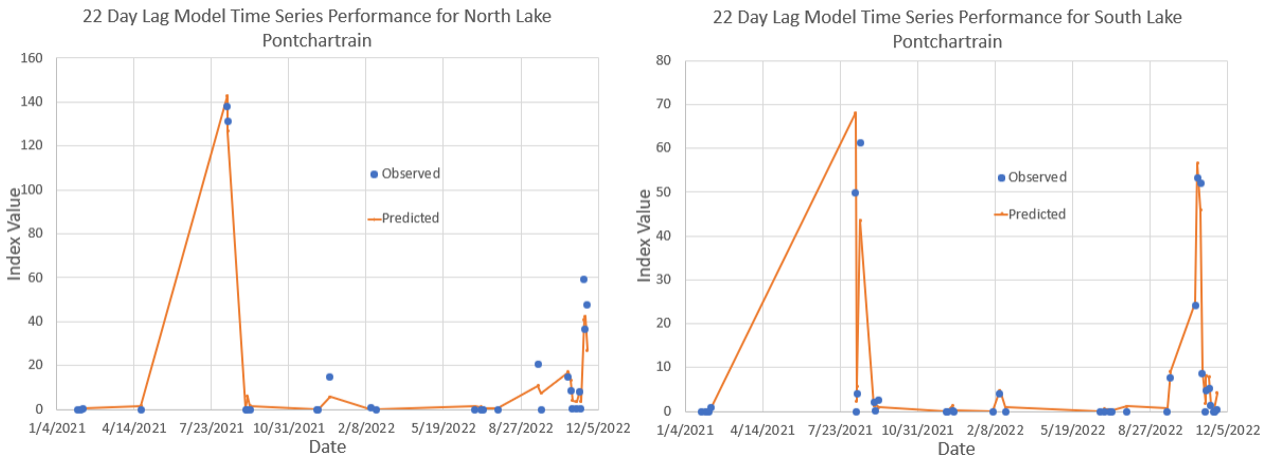


Figure 4.8. Two time series graphs for the 22 day lag model. the left graph representing the North of the lake, and the right representing the South of the lake

Overall, all the graphs show a general trend of a bloom in the summer of 2021 and a bloom in fall of 2022. Generally it can be seen that all the models predict slightly below the observed value with a few instances of over predicting. These instances of over prediction can be seen in the early fall around September 2021 and in the late summer around August 2022. These minor peaks can be observed to some degree in all the models except the 22 day model, Figure 4.8. These graphs show that temporally the models perform well as it is not bias to either the spring, summer, fall or winter months. This is an important factor in the model as it is able to successfully predict index values no matter the season.

### 4.3. Model Forecast Spatial Maps

The following section will explore the spatial performance of the models more in-depth than in the previous section. By taking the observed and predicted values from the validation data set the values and by using the kriging tool in ArcMap, generating averaged spatial maps. Below are 14 figures containing a total of 28 maps. Each figure represents a day with known index values and contains a map representing the observed value and another map representing the model generated predicted values. Since data availability limitations exists, the days that could display the most area of the lake where chosen.

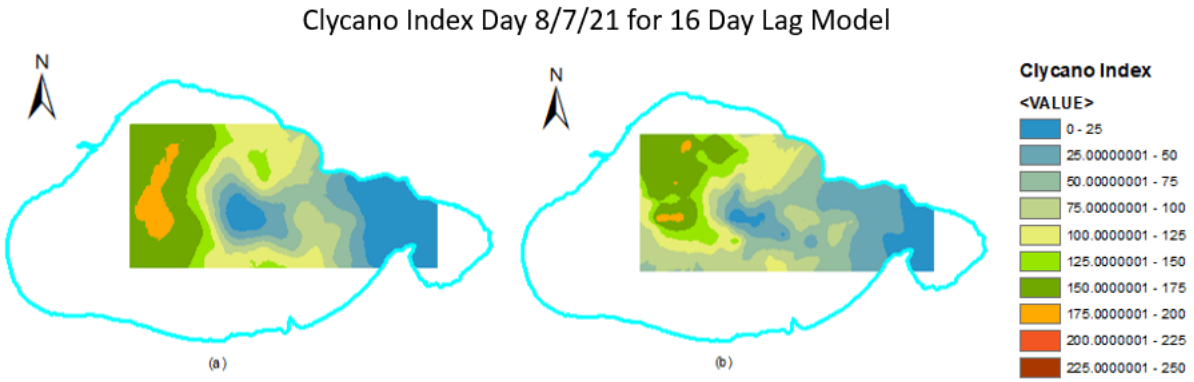


Figure 4.9. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.9 above is the comparison maps of the 16 day lag models predicted values (b) versus the observed values (a) for August 7$^{th}$ 2021. Overall the location of index values of the bloom are well represented by the predicted values. The magnitude of the bloom was under predicted as the orange polygon is significantly smaller in the predicted map. The low index value is well represented in the center-east and to the east of the map.

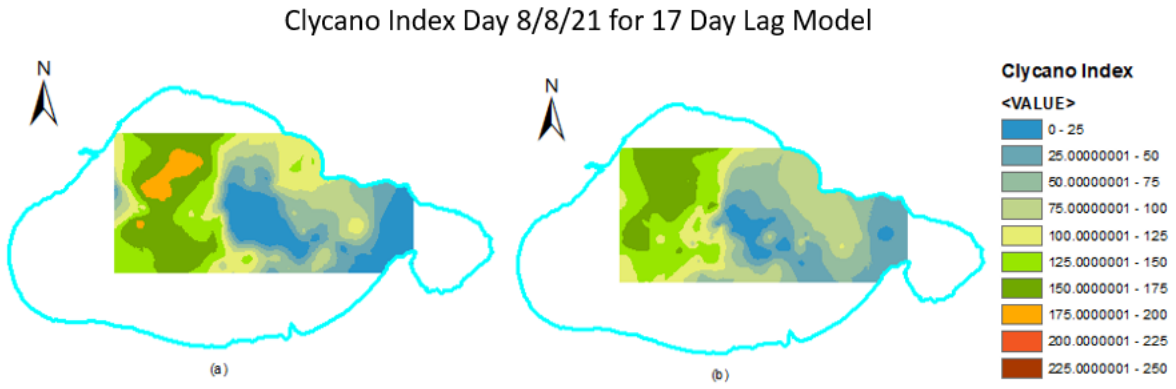


Figure 4.10. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.10 above is the comparison maps of the 17 day lag models predicted values (b) versus the observed values (a) for August 8$^{th}$ 2021. Overall the location of index values of the bloom are well represented by the predicted values and are very similar with Figure 4.9 as it occurs a day later. Like Figure 4.10 the magnitude of the bloom was under predicted as the orange polygon is non-existent in the predicted map. The low index value is well represented in the center-east and to the east of the maps.

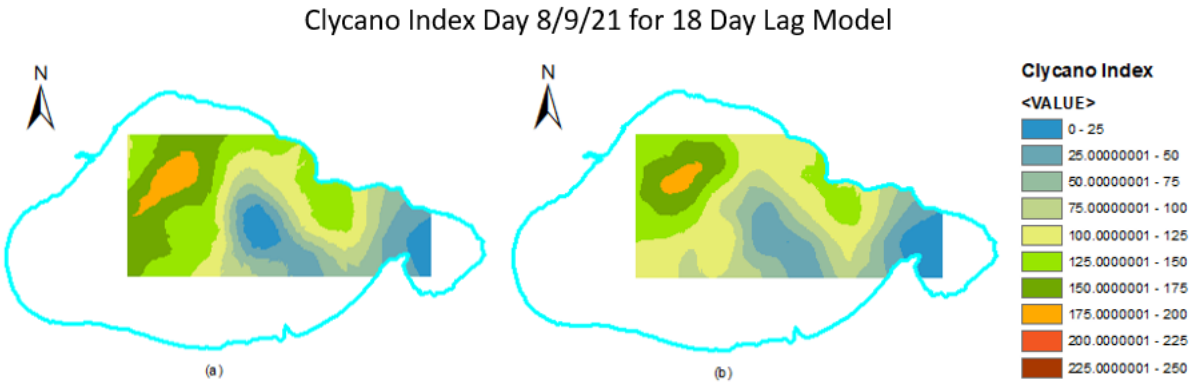


Figure 4.11. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.11 above is the comparison maps of the 18 day lag models predicted values

(b) versus the observed values (a) for August 9th 2021. Overall the location of index values of the bloom are well represented by the predicted values. The magnitude of the bloom was slightly under predicted as the orange and green polygon are smaller in the predicted map versus the observed map. The low index value is well represented in the center-east and to the east of the maps; however unlike in Figures 4.9 & 4.10 the center is slightly over predicted.

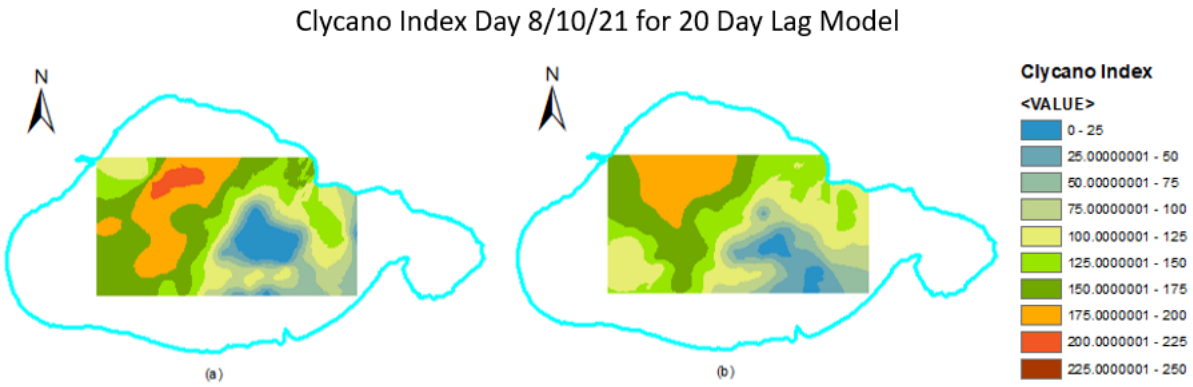


Figure 4.12. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.12 above is the comparison maps of the 20 day lag models predicted values (b) versus the observed values (a) for August 10th 2021. Overall the location of index values of the bloom are well represented by the predicted values. The magnitude of the bloom was slightly under predicted as the red polygon is absent in the predicted map, and the orange polygon does not extend as far south as in the observed map. The low index value is well represented in the center-east of the maps; but, like in Figures 4.11 the center is slightly over predicted as the dark blue polygon is smaller.

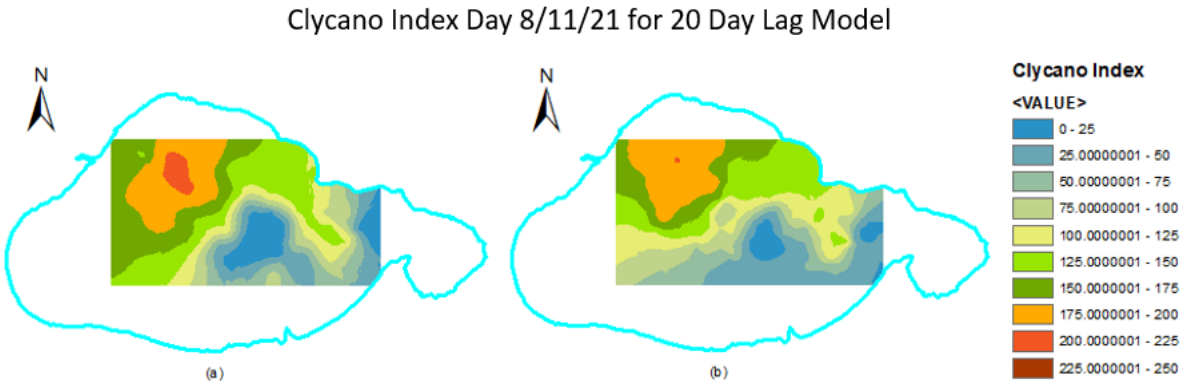


Figure 4.13. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.13 above is the comparison maps of the 20 day lag models predicted values (b) versus the observed values (a) for August 11$^{th}$ 2021. Overall the location of index values of the bloom are well represented by the predicted values. The magnitude of the bloom was slightly under predicted as the red polygon is smaller in the predicted map. The low index value is well represented in the center-east of the graph; also, like in Figure 4.12 the center-east is slightly over predicted as the dark blue polygon is smaller.

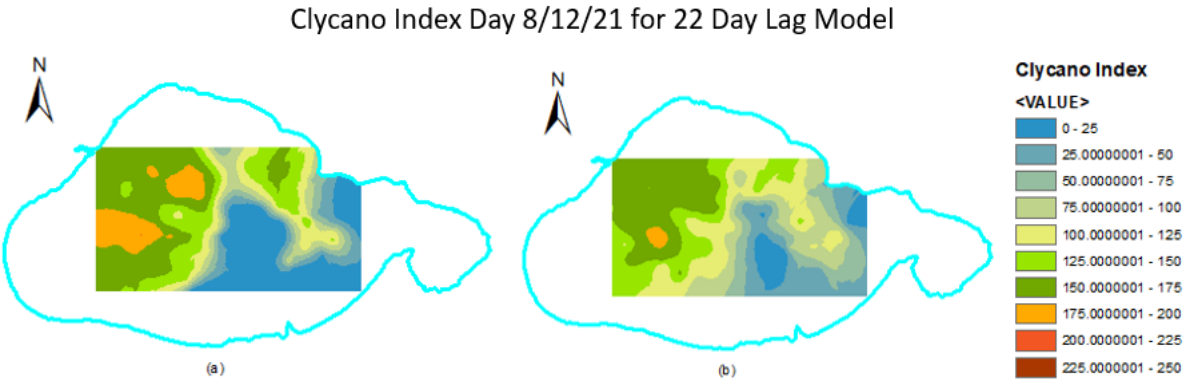


Figure 4.14. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.14 above is the comparison maps of the 22 day lag models predicted val-

ues (b) versus the observed values (a) for August 12th 2021. Overall the location of index values of the bloom are well represented by the predicted values. The magnitude of the bloom was slightly under predicted as the orange and green polygon are smaller in the predicted map versus the observed map. The low index value is well represented in the center-east and to the right of the maps. Like in Figure 4.13 the center-east portion is slightly over predicted as the dark blue polygon is smaller, and the yellowish polygons are bigger.

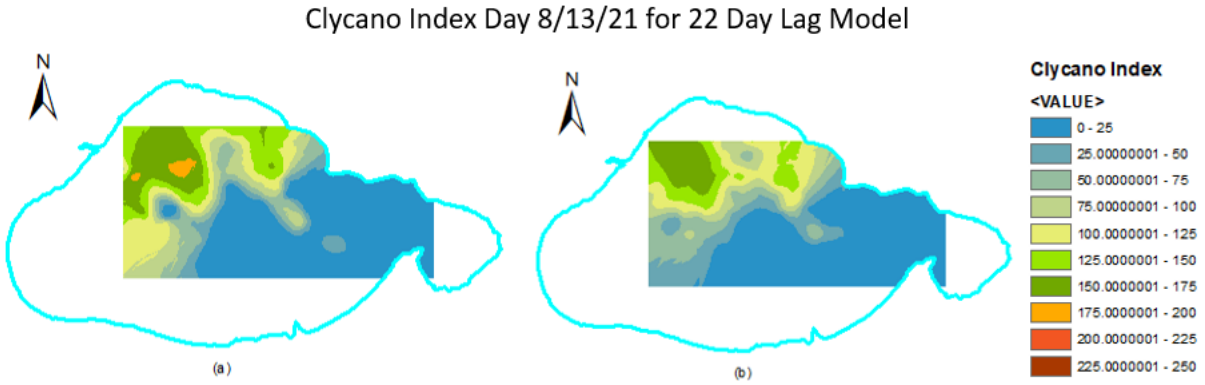


Figure 4.15. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.15 above is the comparison maps of the 22 day lag models predicted values (b) versus the observed values (a) for August 13th 2021. Overall the location of index values of the bloom are well represented by the predicted values. The magnitude of the bloom was slightly under predicted as the small orange polygon is non-existent in the predicted map. The low index value is well represented in the center-east and to the right of the graph, and is not over predicted like in previously discussed maps.

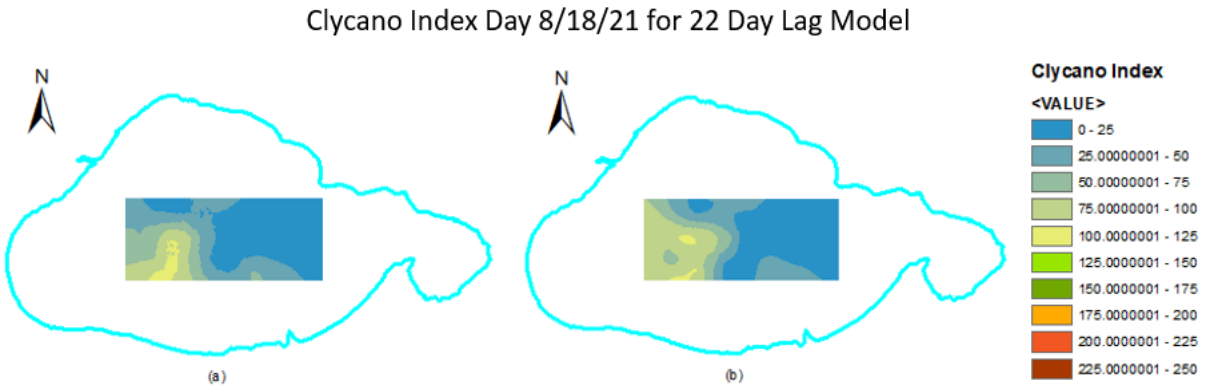


Figure 4.16. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.16 above is the comparison maps of the 22 day lag models predicted values (b) versus the observed values (a) for August 18$^{th}$ 2021. Overall the location of index values of the bloom are well represented by the predicted values; however, the predicted map shows the bloom to be more west and the observed map show it to be more south-west.

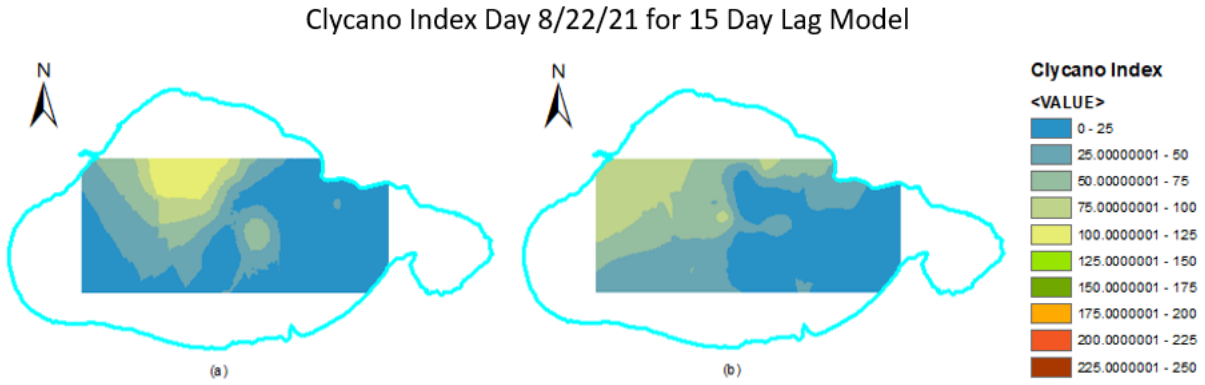


Figure 4.17. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.17 above is the comparison maps of the 15 day lag models predicted values (b) versus the observed values (a) for August 22$^{th}$ 2021. The location of index values of the bloom, while not terrible, are not as well represented as in other comparison

maps. The bloom in the north is not well represented as the observed yellow polygon is more center-north while the yellow polygon is missing in the observed map and the bloom is shown to be more north-west. Overall this is the most poorly predicted day out of the maps chosen.

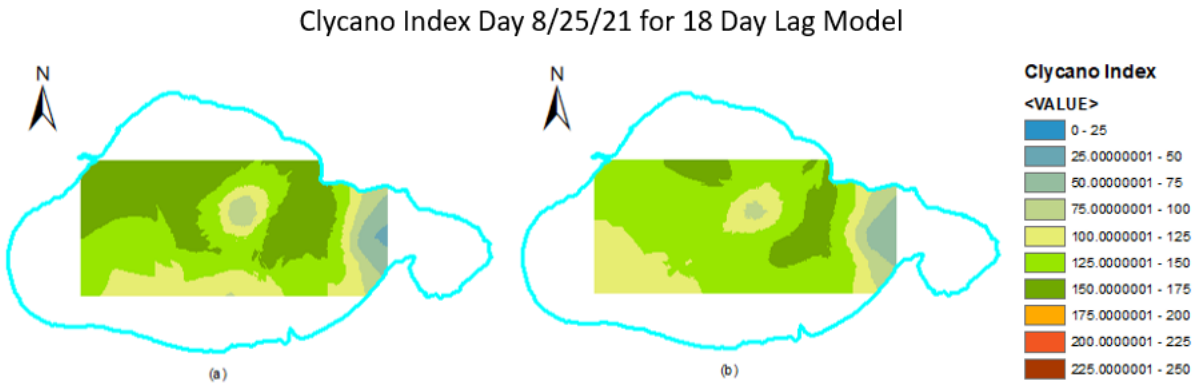


Figure 4.18. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.18 above is the comparison maps of the 18 day lag models predicted values (b) versus the observed values (a) for August 25$^{th}$ 2021. Overall the location of index values of the bloom are decently represented by the predicted values. The predicted map does a good job of showing the the low index values in the center of the map, but under predicts the index values in the north of the map.

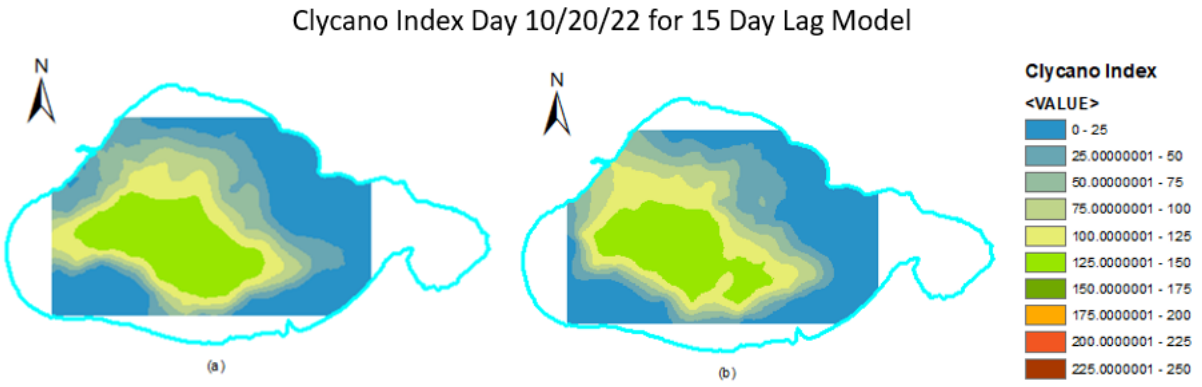


Figure 4.19. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.19 above is the comparison maps of the 15 day lag models predicted values (b) versus the observed values (a) for October 20$^{th}$ 2022. The location of index values of the bloom are represented by the predicted values extremely well. The only fault being that the predicted map seems to be ever so slightly under-predicted

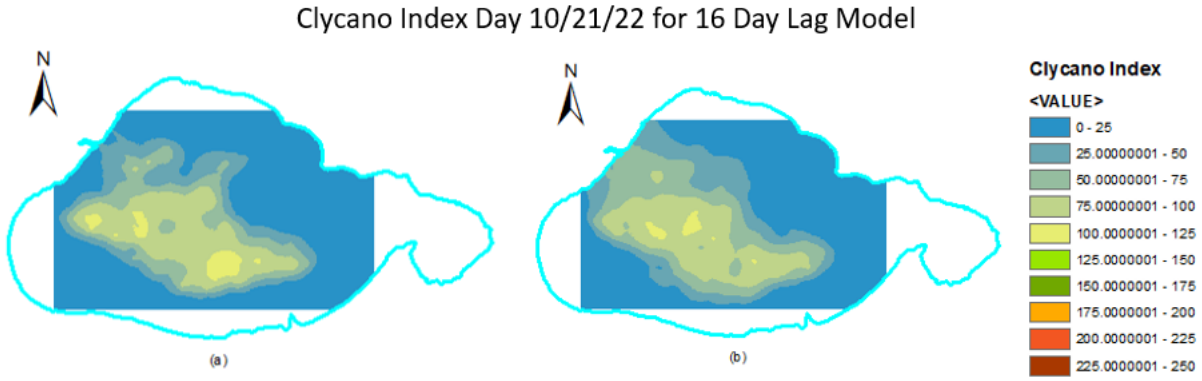


Figure 4.20. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.20 above is the comparison maps of the 15 day lag models predicted values (b) versus the observed values (a) for October 21$^{th}$ 2022. The small yellow polygon in the south is predicted much lower in the predicted map, and the north-west area is slightly

over predicted. Overall, the location of index values of the bloom are represented by the predicted values well.

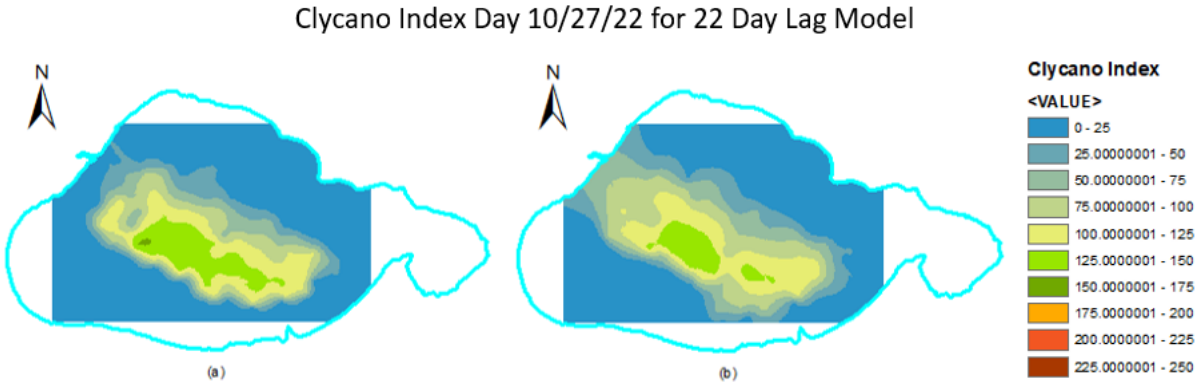


Figure 4.21. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.21 above is the comparison maps of the 22 day lag models predicted values (b) versus the observed values (a) for October 27$^{th}$ 2022. Overall, the location of index values of the bloom are represented by the predicted values well; although, the index value for the predicted map is slightly under predicted.

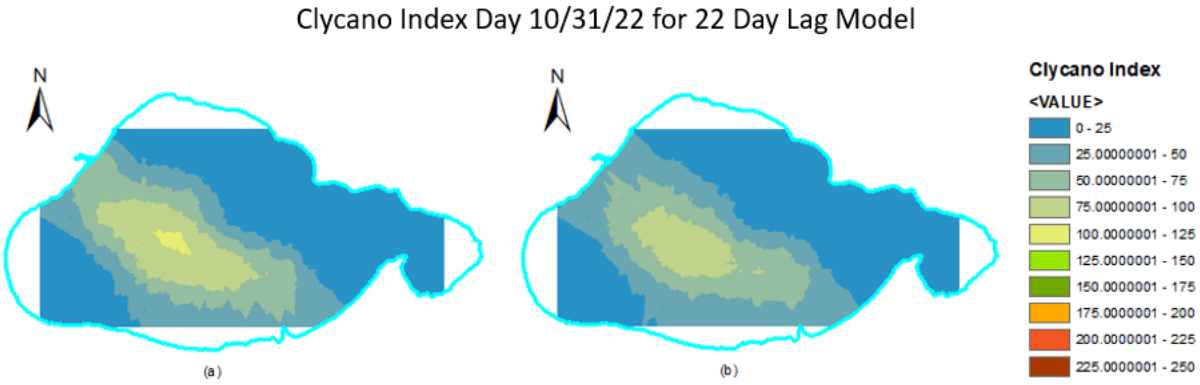


Figure 4.22. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

Figure 4.22 above is the comparison maps of the 22 day lag models predicted val-

ues (b) versus the observed values (a) for October 31$^{th}$ 2022. Overall, the location of index values of the bloom are represented by the predicted values well. Like Figure 4.21, the index value for the predicted map is slightly under predicted as the small center yellow polygon is absent in the predicted value

After reviewing all the spatial map comparisons it can be confirmed that the statistical and temporal assessment was correct, and the ability to confidently produce forecasted spatial maps of CyanoHABs using MODIS-Aqua data is possible for Lake Pontchartrain. Even the worst statistically performing model for the validation data set, the 20 day lag model, produced competent maps for August 10$^{th}$ (Figure 4.12) and August 11$^{th}$ (Figure 4.13). Figures 4.9-4.16 are a good example of the use of 5 of the 8 models to accurately predict a massive bloom in Lake Pontchartrain, and visualize it dissipate over time. The ability to use several different models to accurately predict a bloom over a long period such as a week could prove to be valuable in warning the public and researching mitigation solutions. Having so many viable models also helps with the issue of data availability, as satellite data can be spotty.

### 4.4. Validation With Independent Data Set

Independent data from 2019 that was not used in developing the models was used to independently validate the performance of the models. Specifically MODIS-Aqua data from June 2019 to August 2019 was used to predict CIVs from July 2019 to September 2019 as these dates displayed the most CyanoHABs by the NCCOS Clycano monitoring system. Statistically the 21 day model performed the best out of the 8 producing a correlation coefficient of .2145, a MAE of 32.7545, and a RMSE of 43.1433, shown by Table

4.3. These statistics are nowhere near the performance of the data from 2021 and 2022. However while not great, the MAE of 32.7545 and RMSE of 43.1433 is not totally unacceptable considering the range of the predicted data is 0-250.

Table 4.3. Model Performance Metrics of 21 Day for 2019

| Metric | Value |
|---|---|
| $R^2$ | 0.2145 |
| MAE | 32.7545 |
| RMSE | 43.1433 |
| Instances | 3295 |

Along with producing the best statistical metrics, using the 2019 data the 21 day model produced one of the best temporal graphs and the best spatial map. Figure 4.23 shows the observed versus predicted values for the entire lake from July 4, 2019 to September 22, 2019. As it can be seen the model does not predict zero value (no bloom) days very well in the summer months of June and July and the peak in July was not predicted accurately. However, in September the model predicts zero value days much more closely, except for one day. Figure 4.24 shows the spatial analysis of the 21 day model's prediction of the bloom occurring on August 14, 2019. While not completely displaying the true magnitude of the bloom, the general location of the higher index values and lower index values are displayed competently. While under performing, the results from an independent validation data set are encouraging to the overall future of the model's improvement.

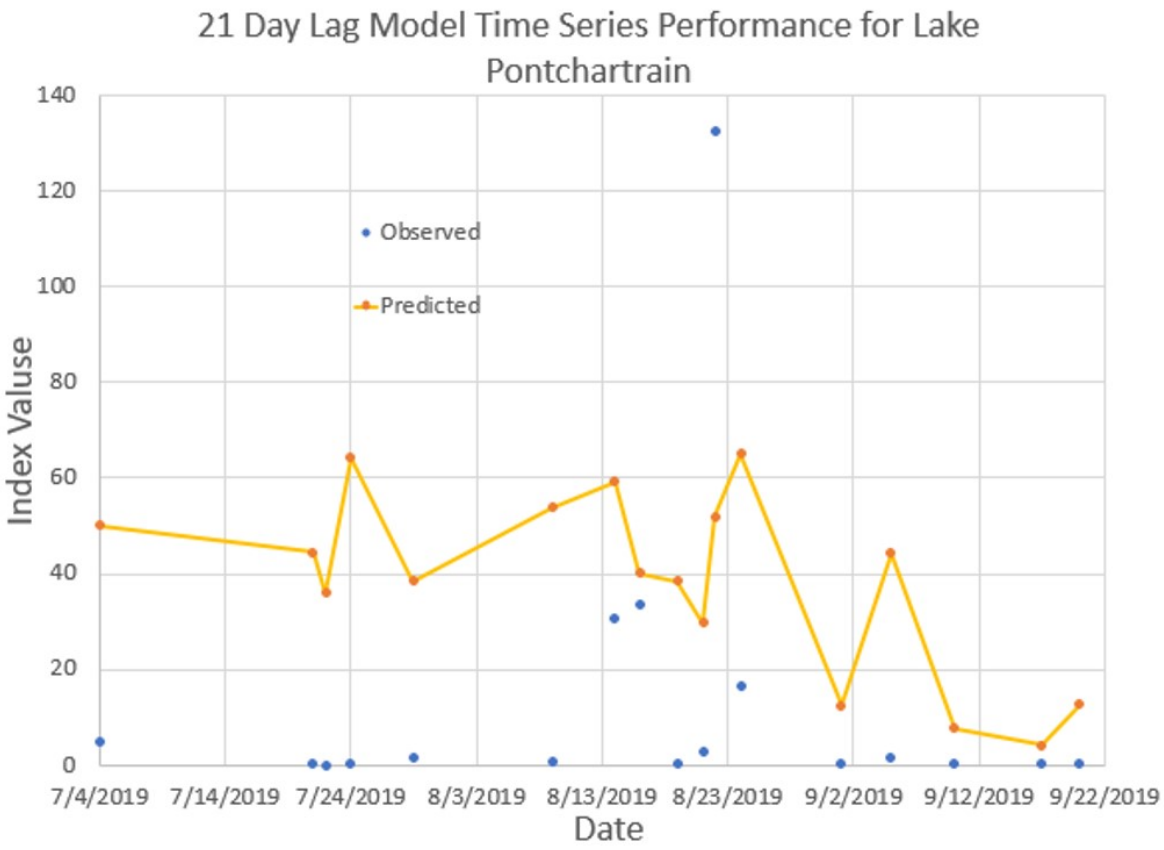


Figure 4.23. 2019 Time Series for the 21 Day Model

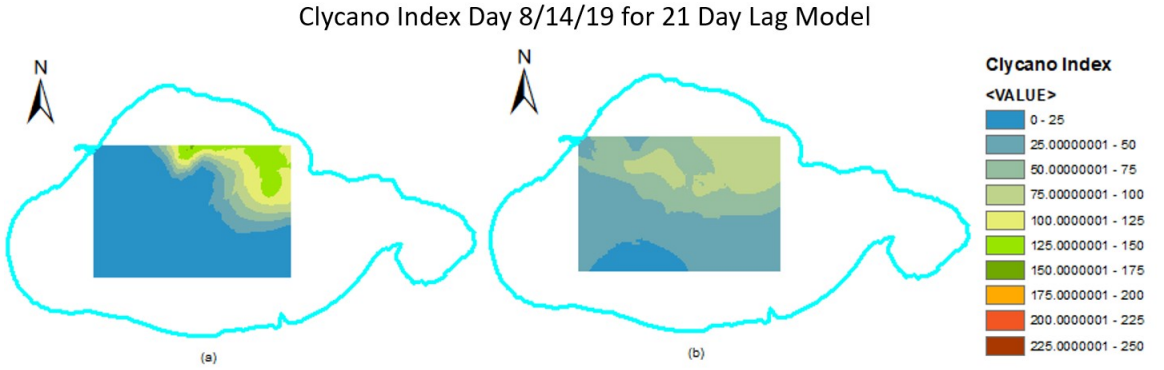


Figure 4.24. Spatial Maps Of Observed Index Value (a) and The Model Predicted Index Value (b).

# Chapter 5. Summaries and Conclusions

Major findings and contributions of this thesis can be summarized as follows

## 5.1. Identification of Important Time Lags and Antecedent Environmental Conditions for CyanoHABs

It was found that the formation of CyanoHABs are controlled by antecedent environmental conditions represented by satellite remote sensing data 15-22 days before a possible CyanoHAB event occurs. These eight time lag days were identified by using MODIS-Aqua satellite remote sensing data from all bands and the software tool WEKA. The remote sensing OC data included the sea surface reflectance bands (Rrs) at the ten different wavelengths of 412nm, 443nm, 469nm, 488nm,531nm, 547nm, 555nm, 645nm, 667nm, and 667nm, the attenuation coefficient (Kd) 490 nm, chlorophyll concentration (chlor-a),particulate organic carbon concentration (poc), instantaneous photosynthetically available radiation (ipar), photosynthetically available radiation (par), normalized florescence height (nflh), the dimensionless aerosol optical thickness at 869nm (aot 869), and the dimensionless aerosol angstrom exponent (angstrom). The reflectance bands represent the physical and biochemical factors that affect the formation, abundance, and persistence of CyanoHABs. In addition to the OC data sea surface tempertature (SST) was downloaded and included as temperature plays a significant contributing role in the development of HABs, and to the identification of the ideal day lag time. While certain variables like Chlorophyll-a, PAR, SST, and NFLH certainly have a proven link to algal blooms, many other variables also affect the occurrence of blooms such as pH, salinity, and nutrients, which can be represented by the remaining remote sensing data. For this reason, all remote sensing variables were downloaded and considered for identifying the optimal day

lag threshold, with the exception of PIC that did not have enough viable data. Prior to the 15-day threshold, the models exhibited suboptimal performance compared to their performance beyond the 15-day mark, as evidenced by a lower correlation coefficient, indicating inadequate predictive capability. After the 22-day threshold it was observed that the day lag model performance of 22 days onward followed a decreasing correlation coefficient trend. Therefore, from the range of lags spanning 30 days, a judicious selection was made to include the eight days (15-22) that demonstrated the most robust correlation with the occurrence of CyanoHABs.

### 5.2. Development of Artificial Intelligence-Based Models for Forecasting Cyano-HABs

Based on these eight time lags, a series of eight forecasting models with the lead-time ranging from 15 – 22 days have been presented in this thesis for issuing early warnings of potential CyanoHABs. The 8 models were developed by using MODIS-Aqua satellite remote sensing data as the model input data and the NCCOS nowcasting Clycano index values as the output data within the software platforms WEKA and R studio. The models can be utilized to predict CyanoHABs in Lake Pontchartrain 15, 16, 17, 18, 19, 20, 21, and 22 days in advance with a model representing each day lag. These models were chosen as they performed the best out of the 30 models developed to give an 8-day forecasting window. While the two algorithms(software tools) XGBoost and Random Forest (RF) performed comparably, the RF algorithm was chosen as it was available to use within the WEKA platform, making manipulation, testing, validation, and saving of many different models easier. The XGBoost model had to be used within R studio and required significantly more user manipulation to function properly and produce results. Through

an iterative process of comparing different RF models with varying parameters the best combination of parameters that produced the highest correlation coefficient was the 50-fold 250-iteration RF model. Statistically all of the 8 models performed well with a correlation coefficient above .90, a MAE below 10, and a RMSE below 20 before validation. After validation the models representing the 15, 16, 17, 18, 19, 21, and 22 day lags all performed with a correlation coefficient above .90, a MAE below 10, and a RMSE below 20. The 20-day lag model produced a correlation coefficient of .8227, a MAE of 12.5987, and a RMSE of 25.3981, which are not as strong as the other 7 but still within an acceptable range for validation. In addition to performing well statistically, the models also performed well temporally and spatially. Temporally, the models predict the seasonal occurrence of CyanoHABs in Lake Pontchartrain and even predicted the uncharacteristically late bloom in October 2022, proving that the models can perform with a wide range of varying seasonally controlled variables. Spatially, the model predicted-values match well with the observed NCCOS Clycano index values. The spatial maps constructed with model predictions show slightly underestimations for some high index value areas and over predicted some low index values. However, the general areas of high index values and low index values are represented very well by the predicted maps. Even the poorest performing statistically validated model, the 20 Day model, performs very well spatially.

These 8 forecasting models are an advancement that currently do not exist for Lake Pontchartrain. The current nowcasting models that give daily information can be useful. However, being able to predict HABs two to three weeks in advance greatly improves the ability of water quality monitoring programs to plan fishing/recreational events in the lake in order to avoid these blooms. Therefore, the eight forecasting models make it possible to

forecast CyanoHABs on a daily basis and thereby inform water quality programs of where and when CyanoHABs are likely to occur so that managers can proactively respond to CyanoHAB events, greatly reducing the CyanoHAB risk to the public health.

### 5.3. Reduction of Cloud Cover Impact on Satellite Remote Sensing Data

A major barrier to the broad application of satellite remote sensing data is the negative impact of cloud cover on remote sensing data. Basically, satellite remote sensing are not available for most days of a year due to the cloud cover, greatly limiting the application of satellite remote sensing data. A major new contribution of this thesis is that the combined application of these 8 models helps reduce the negative impact that cloud cover exerts on remote sensing data. Given the existence of 8 distinct models, the occurrence of cloud cover during the period spanning days 19-22 prior to the desired forecast day does not substantially hinder predictive capabilities. In such instances, if four models are rendered mute there are still four alternative models available for providing reliable predictions for the target day. The capacity to adapt to meteorological conditions, spatial coverage limitations, and the influence of intense glint phenomena on oceanic surfaces significantly broadens the applicability of remote sensing data, thus mitigating the unfavorable implications entailed in its deployment.

### 5.4. Future Works

While the model works well to predict the NCCOS Clycano index values from the training data set years, it does not predict concentration values. In addition the index value range that corresponds to a certain concentration is constantly changing. Being able to predict concentration values with the ever changing index value range could be a

significant improvement to the models. Furthermore, the more information a model possesses the more accurate it will be, and these current models are limited by available data. Combining these model's algorithm with other algorithms that can predict and fill missing data could prove to be highly effective. The independent 2019 validation data set results could be improved by including in-situ data to the training model, such as chlorophyll concentration and wind direction/speed. Along with this in-situ data, studying the effect that a different day lag has on each variable's ability to predict an index value could prove useful. For example, combining a variable that has the higher importance at the 19 day lag interval with a variable that has a higher importance at the 15 day lag interval to make a model that can predict at the 20 day lag interval could produce better results. Overall, the results of this thesis's study are an exciting new development that has much future promise.

# Appendix A. Mat Lab Codes

```
SST=readtable("1_07_22_SST.xlsx",'VariableNamingRule','preserve');
Modis=readtable("1_07_22_Modis.xlsx",'VariableNamingRule','preserve');
tmp=zeros(size(SST,1),1);
for i = 1: size(SST,1)
    tmp(i)=find(ismember(Modis(:,1:2),SST(i,1:2),'rows'));
end
Modis_new=Modis(tmp,:);
Modis_new(:,27)=SST(:,5);
Modis_new(:,[19,24])=[];
Modis_new=renamevars(Modis_new,"Var27","SST");
writetable(Modis_new,'E:\2022 excel data files\Raw MSST files\1_07_22_MSST.csv');
```

Figure A.1. The Matlab code used to match OC data with SST data

Above is the Matlab code used to match the OC data with SST data. The first two lines import the data as a table, and the third line creates an empty matrix to put new data. The following loop finds the rows where the x and y pixels match for the OC data and the SST data, and gives the corresponding row of the MODIS Day. The following line makes a new OC file called Modis_new that pulls the corresponding row from the index file created by the loop. Then the next line puts the corresponding SST file with the Modis_new file. The following lines are clean up where the unnecessary columns are eliminated and columns are renamed correctly. Finally the new file is saved in a place where it can be found easily.

```
SatData=csvread('9_26_21 Modis.csv'); % less points
ClyData=csvread('10_16_21_cly.csv'); % more points

S1=size(SatData,1);
S2=size(ClyData,1);
min_dis=zeros(size(ClyData,1),1);
min_num=zeros(size(ClyData,1),1);
for i = 1 : S2
    tmp_dis=zeros(S1,1);
    for j = 1 : S1
        tmp_dis(j)=norm(SatData(j,2:3)-ClyData(i,2:3));
    end
    [min_dis(i),min_num(i)]=min(tmp_dis);
end
color_val=zeros(size(SatData,1),1);
for i = 1 : S1
    cor_num=find(min_num==i);
    color_val(i)=mean(ClyData(cor_num,4));
end
cv=table(color_val);
SatData1=readtable('9_26_21_MSST.csv','VariableNamingRule','preserve');
SatData1(:,31)=cv(:,1);
SatData1(:,["Pixel-X","Pixel-Y","sstref","bias_sst","stdv_sst"])=[];
SatData1=renamevars(SatData1,"Var31","cly_obs");
writetable(SatData1,'E:\2021_2022 combined\Validation files\Raw with lat long\20day_11.csv');
```

Figure A.2. The Matlab code that matches the CIV data with the MODIS-Aqua OC data

Figure A.2 above is a code that re-samples the CIV data to match the MODIS-Aqua OC data. The first two lines import the data corresponding to desired day lag. The next four lines size the previously imported data sets and prepare empty matrices to be run by the next two loops. The first loop has a second loop within it that finds the CIV data that falls within the MODIS-Aqua OC data and then the outer loop assigns the data to the previously made empty matrices. The next line creates another empty matrix for the mean CIV value to go. The following loop creates the mean CIV value corresponding to every MODIS-Aqua OC row data. The next line creates a table for the CIV data so it has a header, and the next line imports the corresponding MSST data as a table which is the data file created that has both the OC data and SST data. The next three lines add

the CIV values to the appropriate OC/SST data and unnecessary data is deleted. Finally the data is saved with appropriate title denoting what the day lag is (the first number) and what data was used (the last number).

# Appendix B. Weka Model User Guide

WEKA is a user friendly machine learning platform used to develop, test, and validate the models in this thesis. Below is a step by step user guide to describe the process used in this thesis.

## B.1. Step 1: Open WEKA

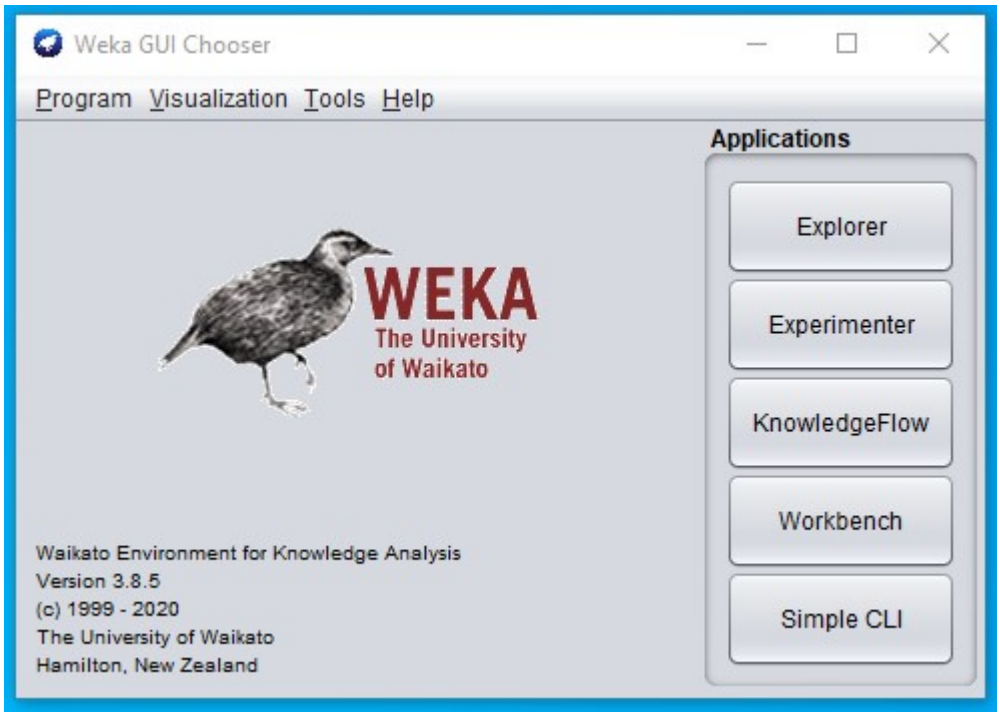


Figure B.1. The opening window when the WEKA program is opened

Figure B.1 above is a picture of the GUI chooser generated when WEKA is opened. There are five applications options to choose from. To continue click on the Explorer application button.

## B.2. Step 2: Open Preprocessing File

To train the models the appropriate day lag file must be imported into the WEKA application by clicking on the open file button in the top most corner of the screen displayed by Figures (B.2 and B.3) below. WEKA can read many forms of files; however, the easiest form to use for this thesis was a csv file as all of the raw data was processed using Excel. Once loaded, shown by Figure B.3, the data can be explored as the attributes are shown as well as some statistics, and the other tabs in the Explorer application become available.

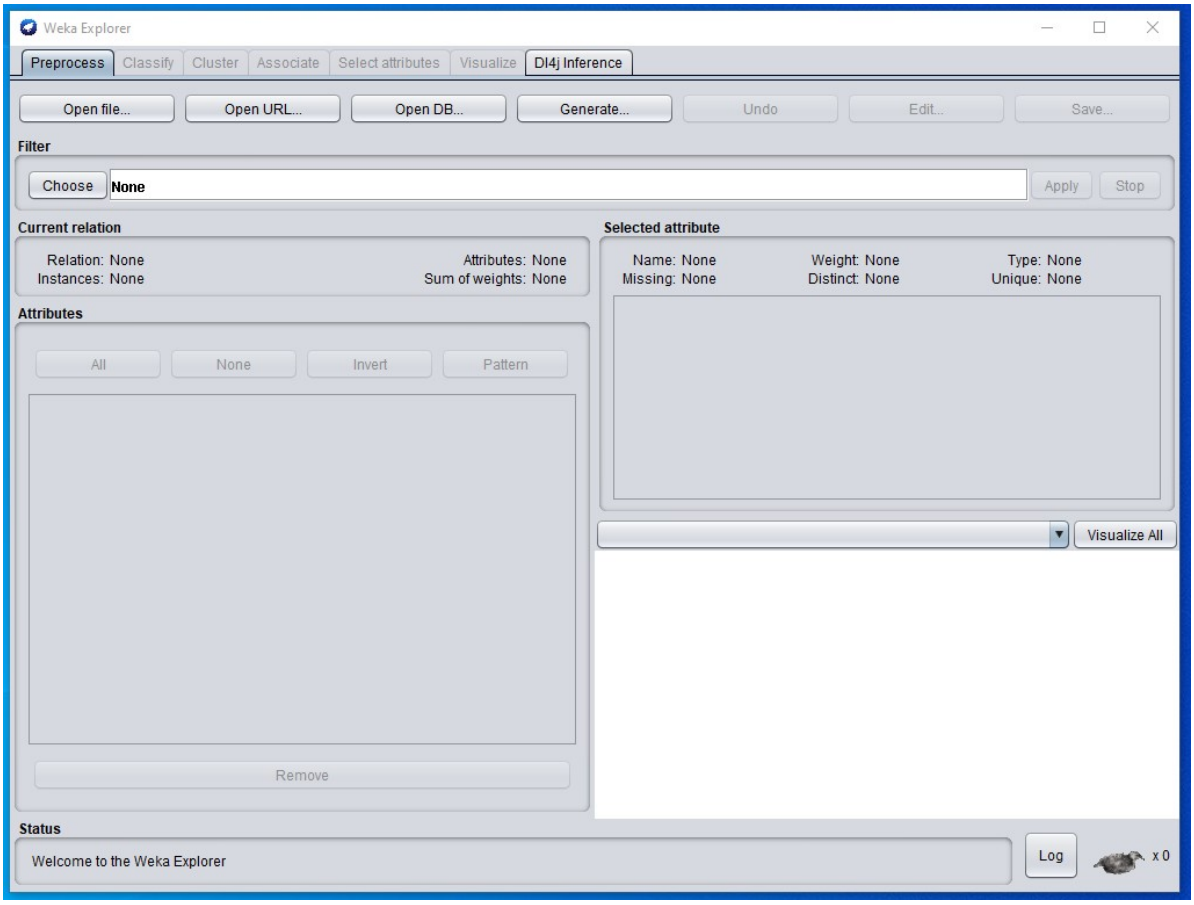


Figure B.2. A picture of the explorer pre-processing tab

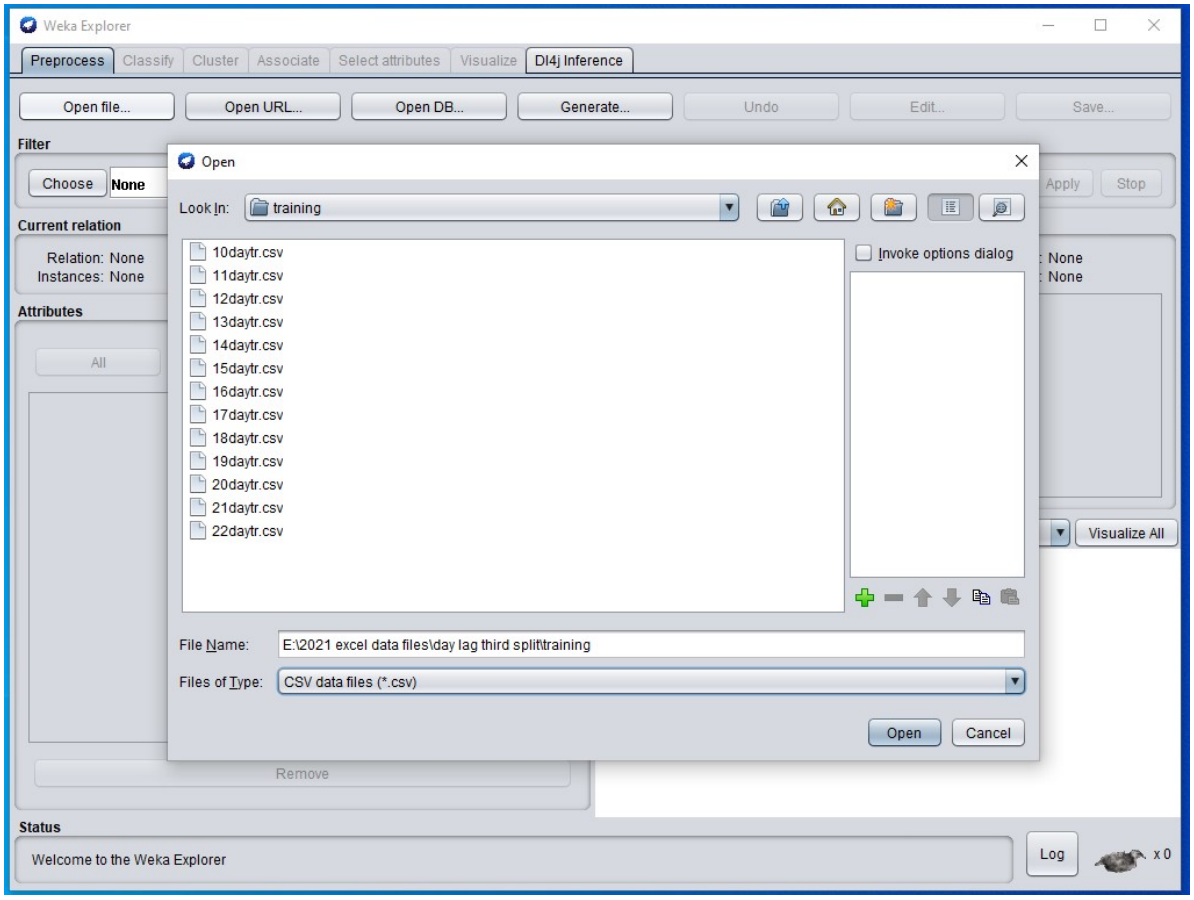


Figure B.3. A picture of the explorer pre-processing tab when open file is clicked

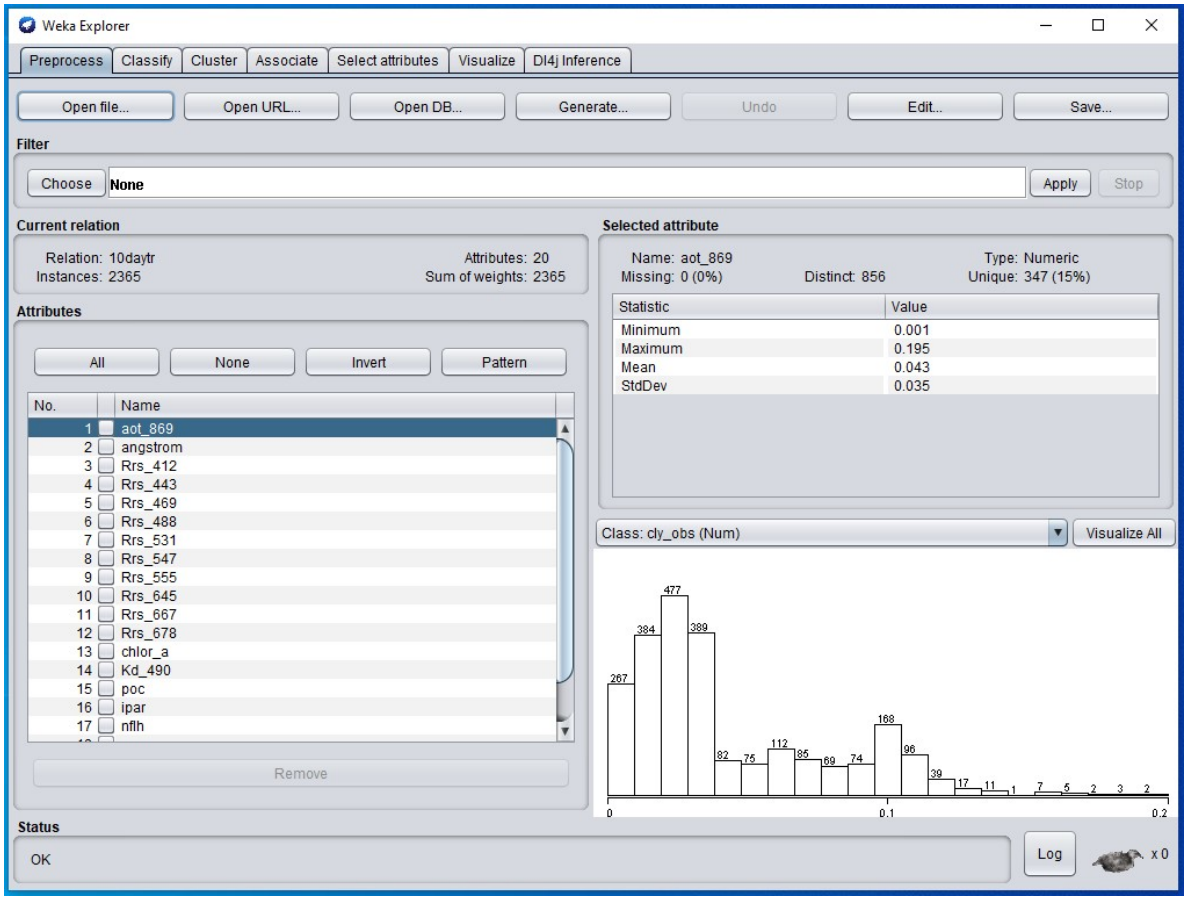


Figure B.4. A picture of what the pre-process data looks like once loaded

## B.3. Step 3: Choose Algorithm and Change Desired Parameters

The next step is to click on the classify tab that is now available and that will pull up a screen shown by Figure B.5. From here the desired algorithm must be chosen. This is done by clicking the choose button in the top left of the screen, and this will display the available algorithms and a brief description shown by Figure B.6. Being an open-source and evolving platform, one can download different algorithms developed by other users that are not included in the programs base download. For this thesis the Random Forest Algorithm was chosen as the final algorithm. Once the algorithm is chosen, the parameters can be changed, which can be done two ways. The first is by right clicking into

the white bar and choosing the enter or edit configuration option where it is possible to manually change the parameters if the corresponding parameter denotation is known. The second is by left clicking into the bar which pulls up the properties tab shown by Figure B.7. It is here that the options can be changed without knowing the parameter denotation. However, research on what each of the categories stands for is needed; which can be found by either googling what each parameter stands for or using the WEKA guide. Also, in the classify tab the different testing options can be seen in the test options section. For this thesis the cross-validation option was chosen with 50 folds.

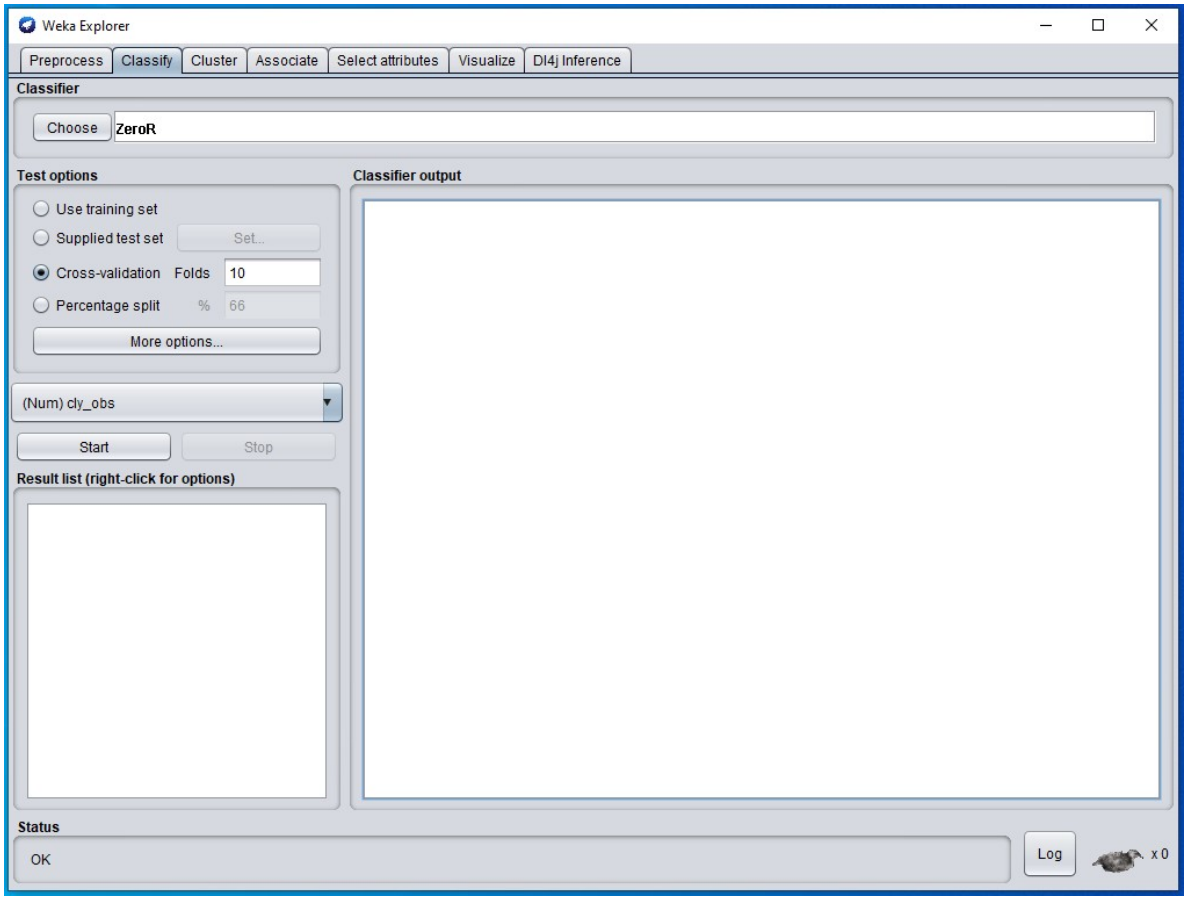


Figure B.5. A picture of the classify tab

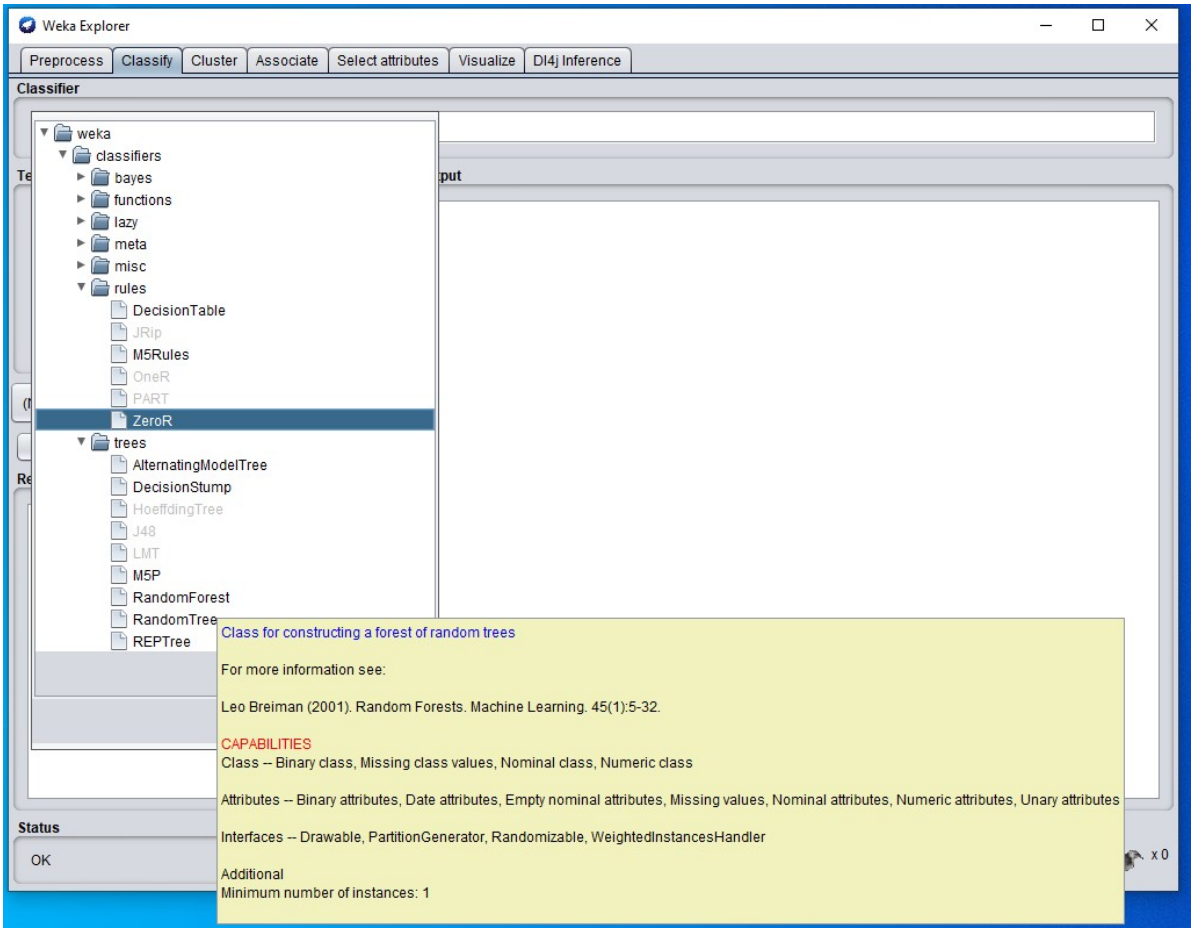


Figure B.6. A picture of the tab showing the different algorithm available

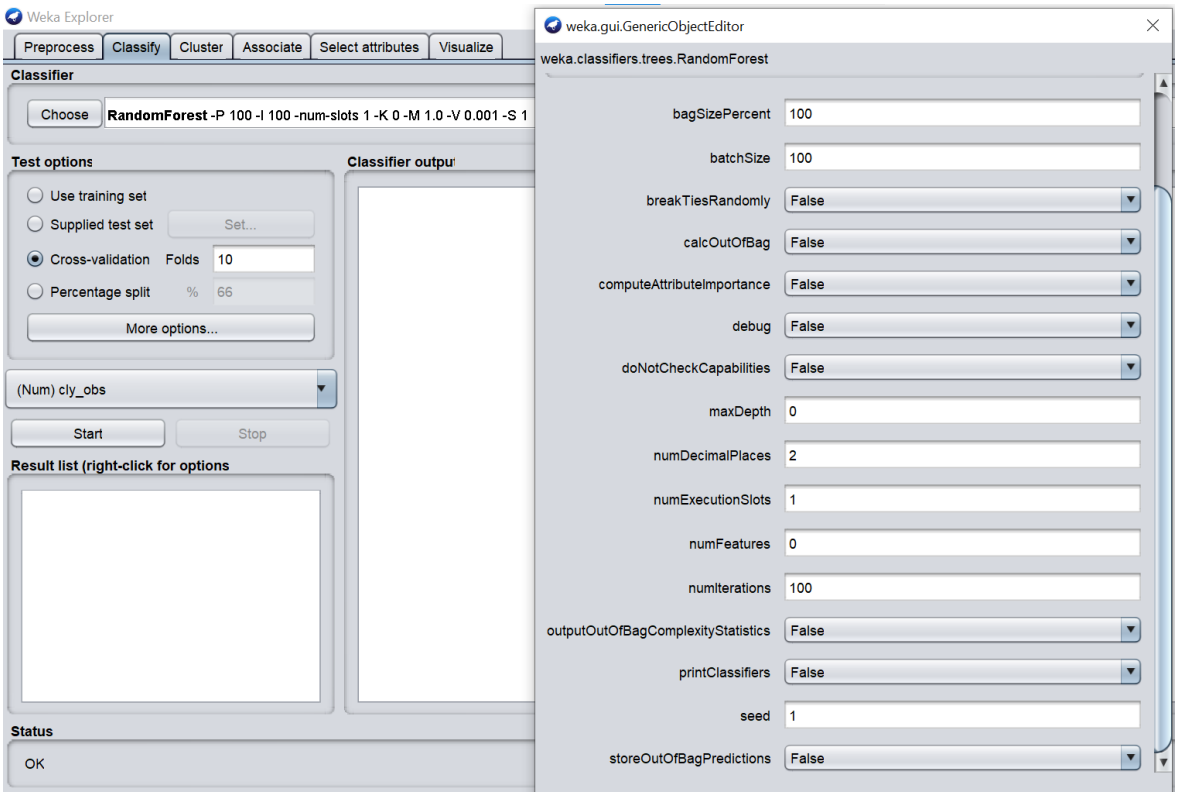


Figure B.7. A picture of the parameters tab for the Random Forest algorithm

## B.4. Step 4: Train and Save the Model

Next the model can be created and trained. Before the creation and training is started, it is pertinent that the correct variable is chosen so that the model predicts the desired outcome. WEKA defaults in choosing the last column in the data file imported as the predicted variable, but it can changed by the pull down tab below the test options section. For this thesis the cly_obs variable was chosen. To run the algorithm one simply clicks the start button located in the middle left of the screen. This will generate a statistical summary in the large white space on the screen, and it is at this stage that the statistical matrix can be recorded as well as variable importance values for each model. The model can be saved by right clicking on the Results option that is generated in the smaller

white box in the lower right corner and clicking on save model. Figure B.8 below depicts the statistical summary and the save model tab.

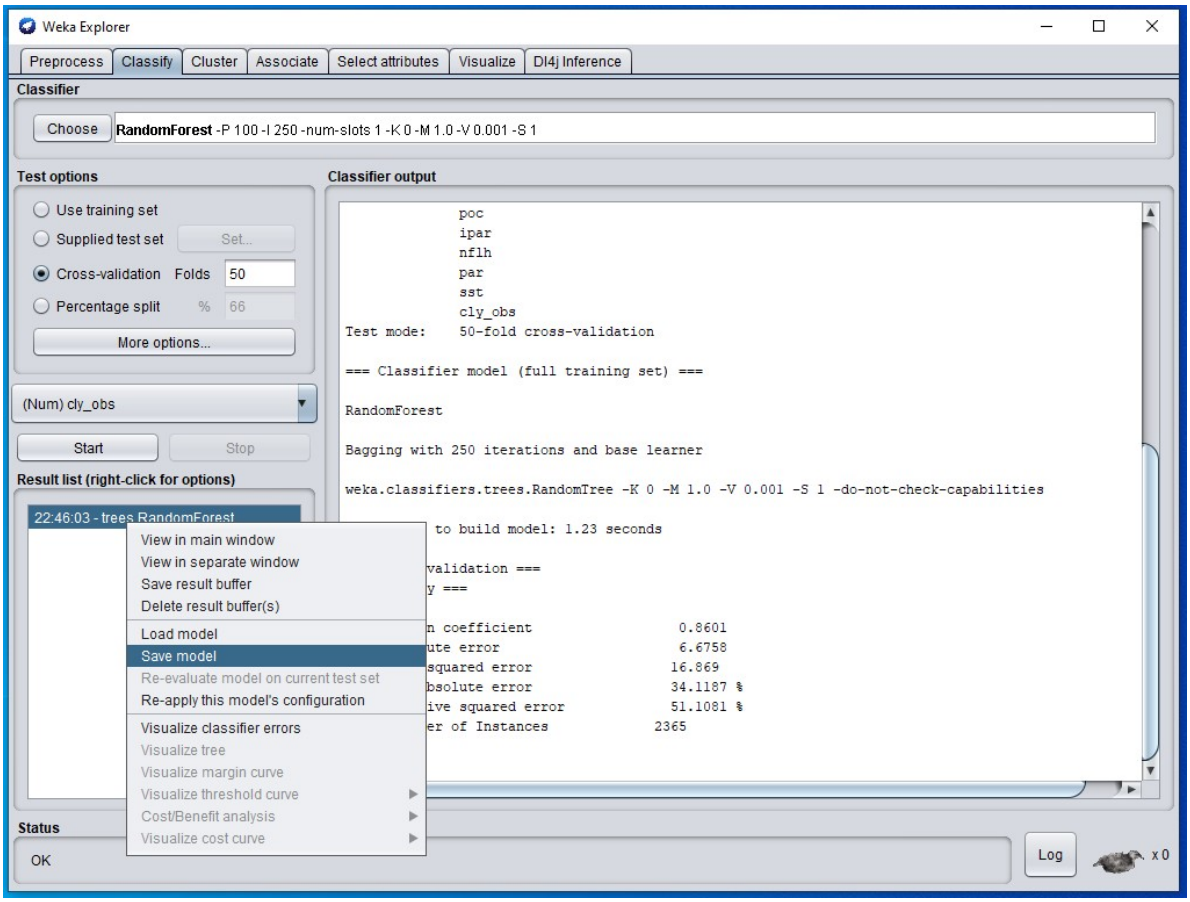


Figure B.8. A picture of the statistical summary and save model tab

### B.5. Step 5: Load and Validate Model

Next the models need to be validated with a supplied test set. If the user hasn't closed out the program, the model won't have to be re-loaded. If the user has closed the program the model will have to be re-loaded, which can be done by right clicking into the small right box and selecting load model, shown by Figure B.9. Once the model is loaded the validating data set must be loaded by selecting the supplied test set option in the test options section and clicking on the set button next to it. This will open the tab labeled test instances, and the open file button needs to be selected which will open a file explorer tab allowing for the selection of the desired validation data set. This process is demonstrated by Figure B.10. The next step, shown by Figure B.11, is to select the proper display of the statistical summary. Clicking on the more options tab will open the classifier evaluation options tab where one can select the desired information to be displayed in the summary section. Clicking the choose button next to output predictions will give options on how one wants to see or save the models output. By selecting plain text the summary will display all the predicted values with its corresponding actual value and error value in the summary section along with the model's statistical metrics. The next step, shown by Figure B.12, is to re-evaluate the desired model to generate predicted values. By right clicking on the model in the results list it will pull up an option underneath the save model option called re-evaluate model on current test set. Selecting this option the desired validation statistical summary is generated in the classifier output section, shown by Figure B.13, and the information can be recorded for evaluation.

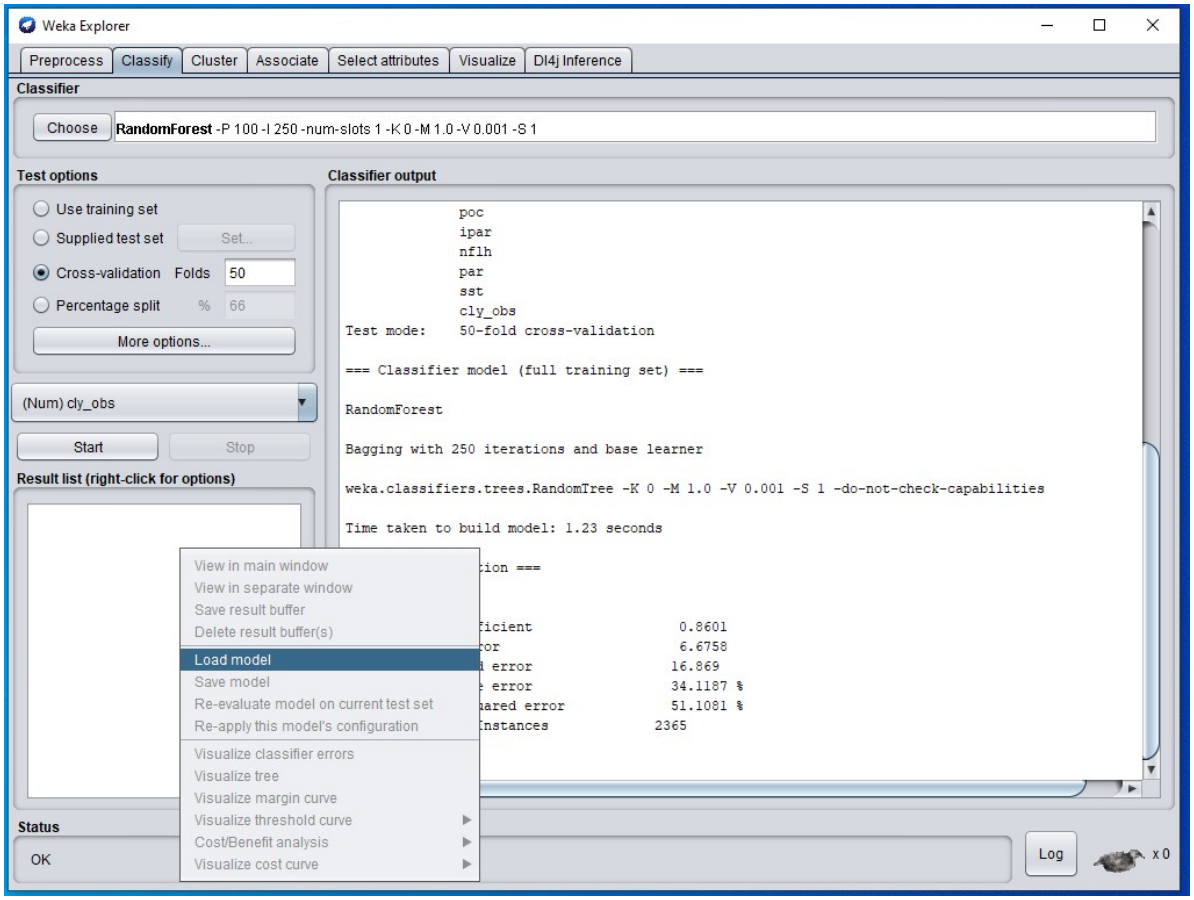


Figure B.9. A picture of the load model option

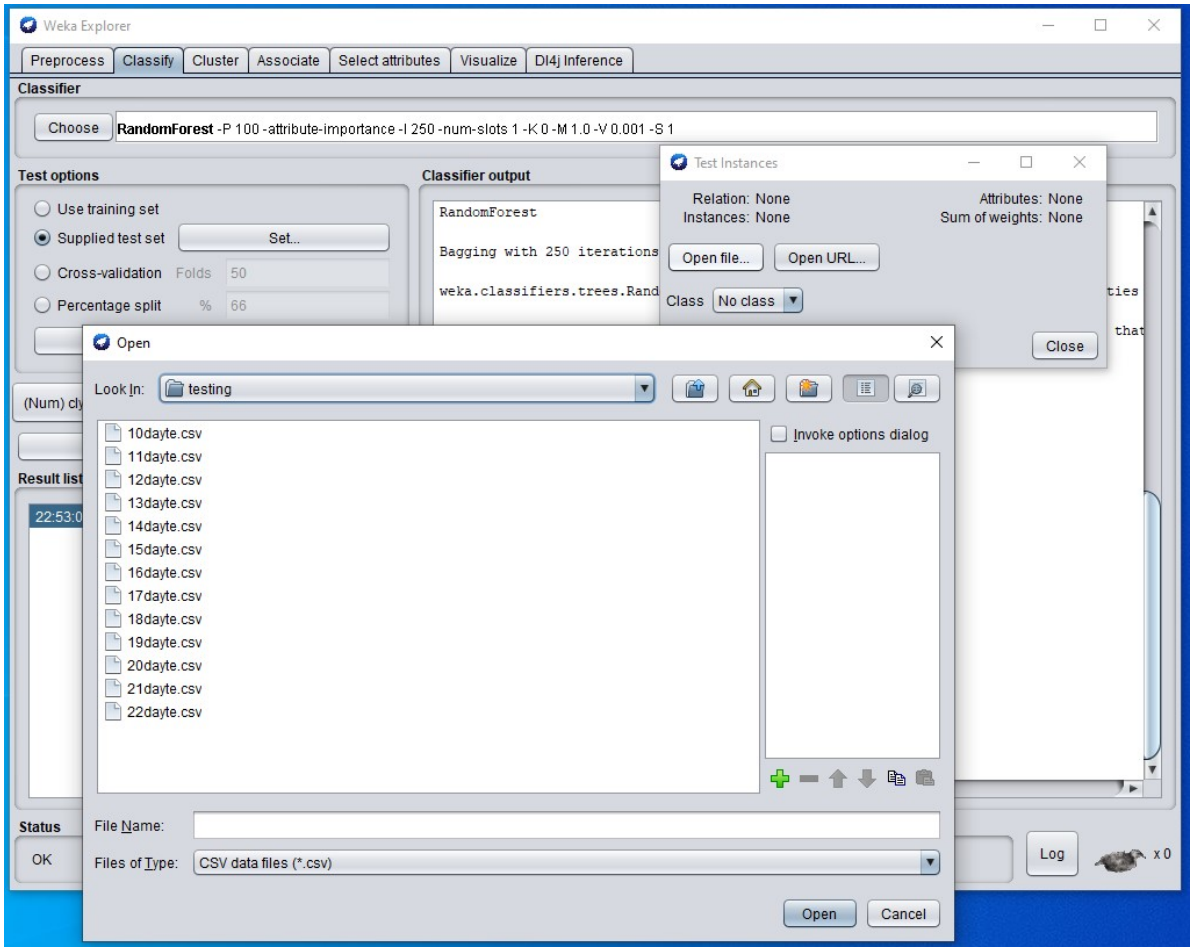


Figure B.10. A picture on how to load a test set

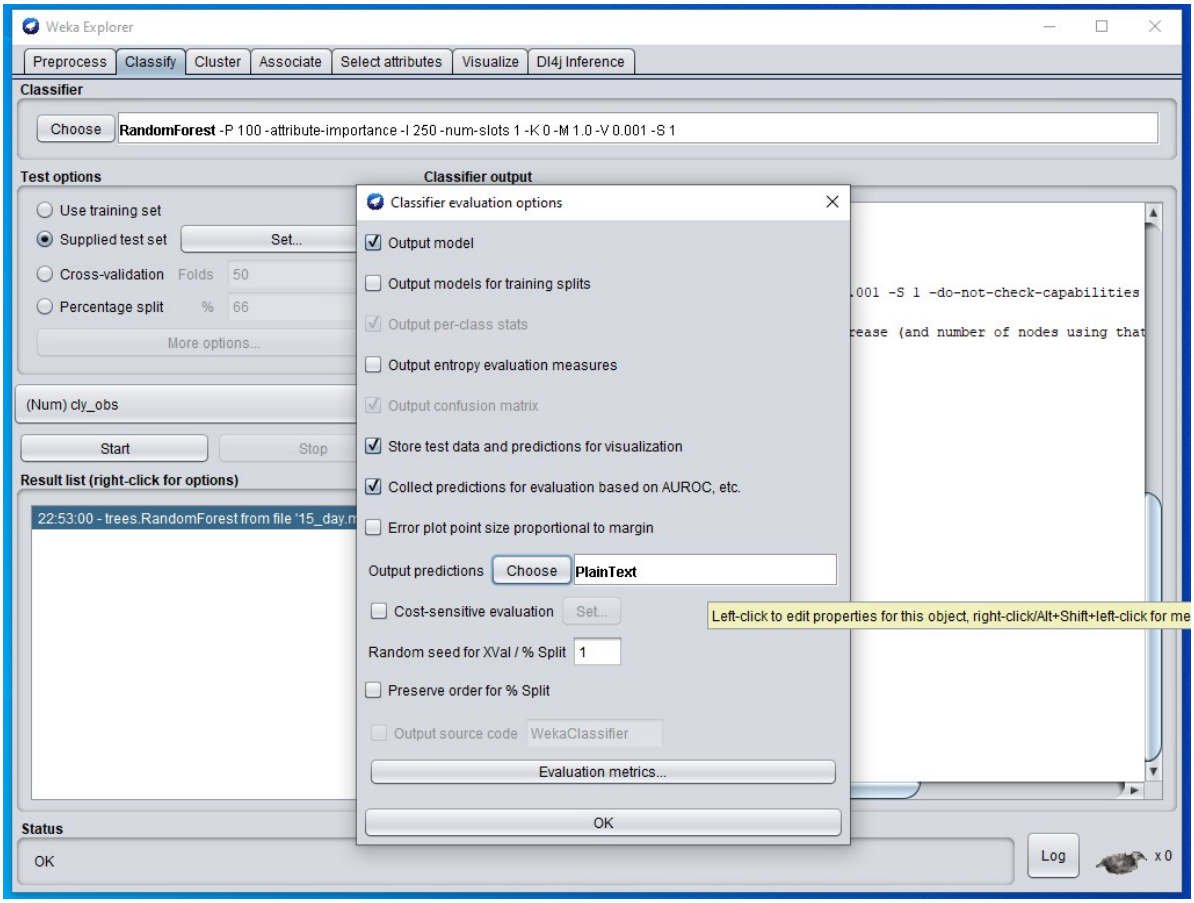


Figure B.11. A picture showing the classifier evaluation options

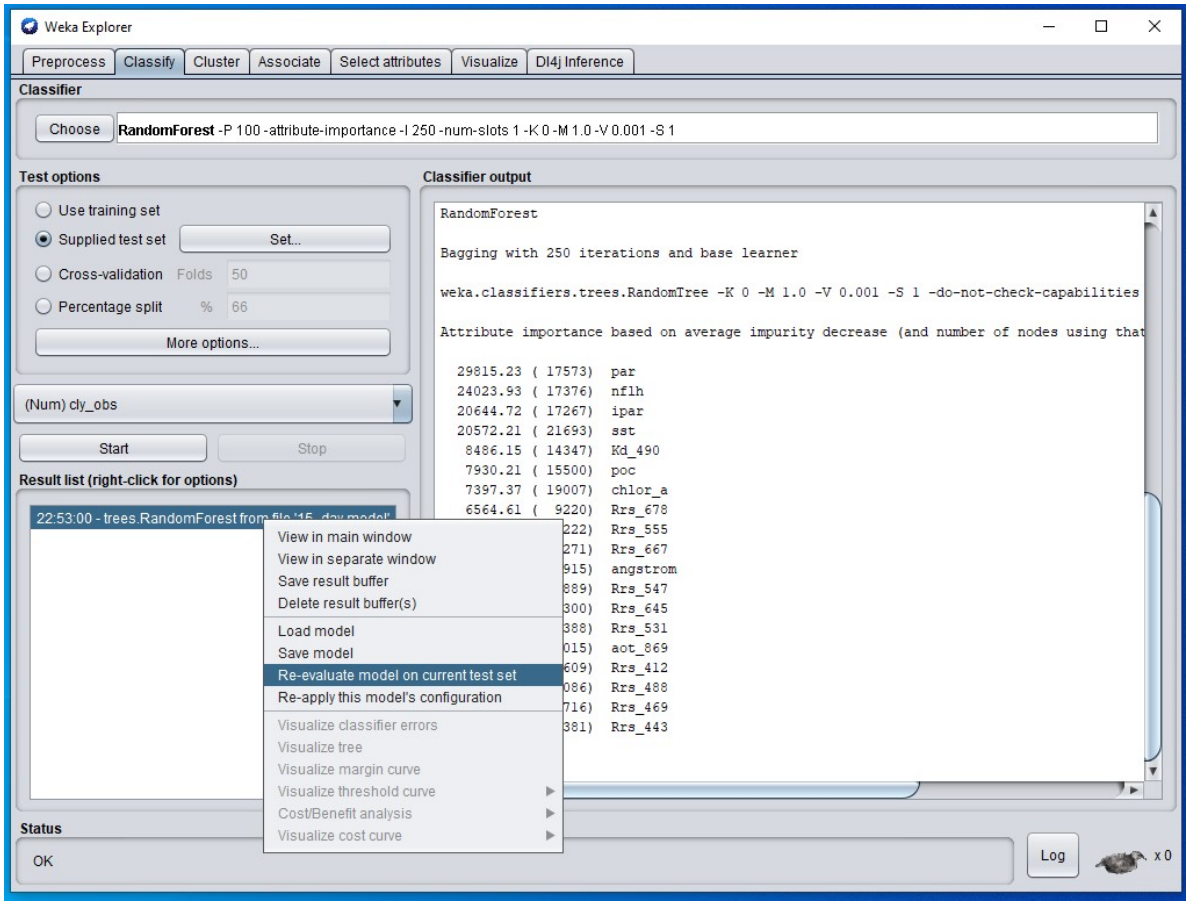


Figure B.12. A picture showing the re-evaluation option of the model

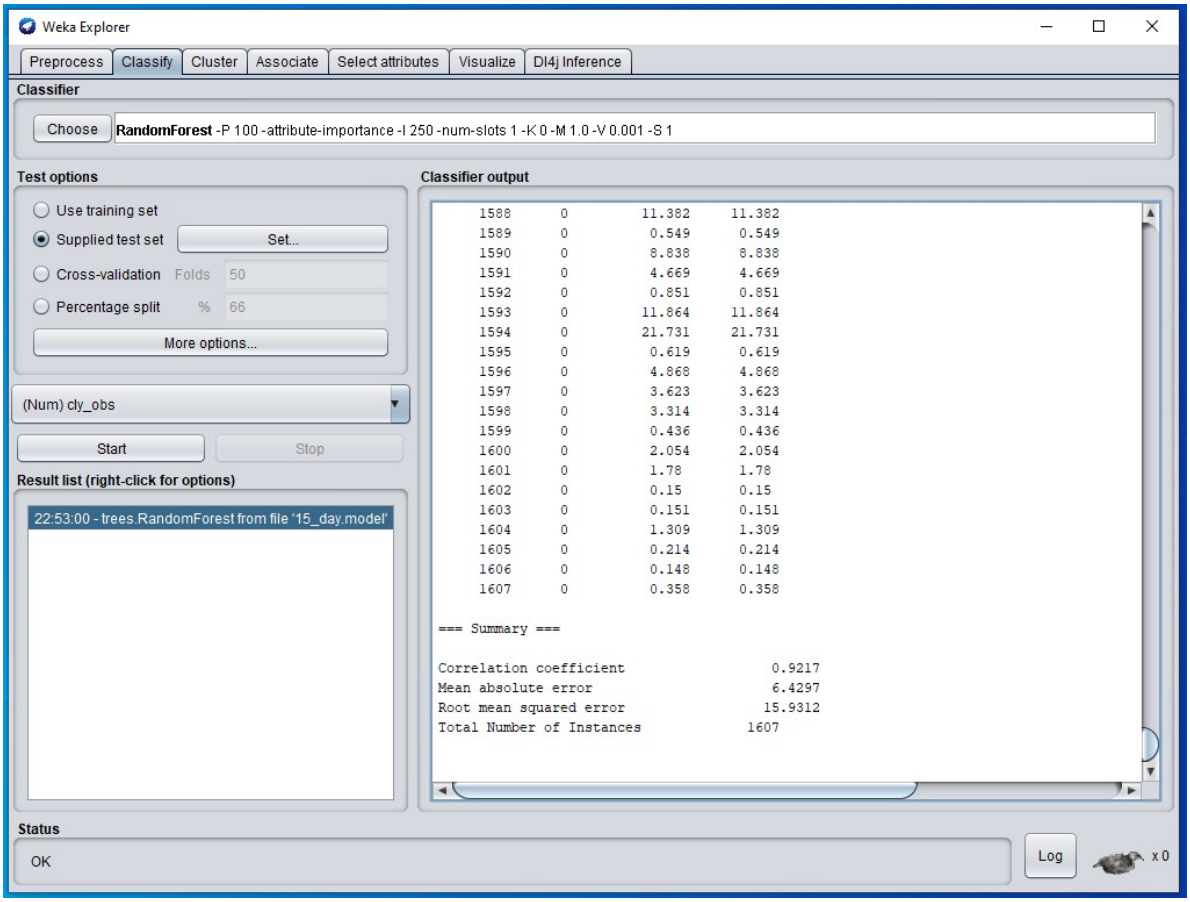


Figure B.13. A picture showing the validation data summary

## Bibliography


[1] Adriance, J., Marx, J., Bourque, C.,and Brit, C. 2018, November. *An overview of Louisiana Department of Wildlife and Fisheries data collected in the vicinity of Lake Pontchartrain and Lake Borgne from 2004 to 2017, as related to the MRGO Rock Dam closure in 2009.* Marine Fisheries Section, Office of Fisheries Louisiana Department of Wildlife and Fisheries.

[2] Ali, J., Khan, R., Ahmad, N.,and Maqsood, I. *International Journal of Computer Science Issues, 9(5)*

[3] Alikas, K., Kratzer, S., Reinart, A., Kauer, T., & Paavel, B. (2015). Robust remote sensing algorithms to derive the diffuse attenuation coefficient for lakes and coastal waters.*Limnology and Oceanography: Methods, 13(8)* , 402-415.

[4] Bailey, S. (n.d.). *Algorithm Descriptions [Review of Algorithm Descriptions].* EarthData. Retrieved April 10, 2023, from https://oceancolor.gsfc.nasa.gov/atbd/

[5] Berdalet, E., Fleming, L. E., Gowen, R., Davidson, K., Hess, P., Backer, L. C., Moore, S. K., Hoagland, P., and Enevoldsen, H. (2016). Marine harmful algal blooms, human health and wellbeing: Challenges and opportunities in the 21st century. *Journal of the Marine Biological Association of the United Kingdom, 96(1).* https://doi.org/10.1017/S0025315415001733

[6] Breiman, L. (2001) "Random Forests." *Machine Learning* vol. 45, no. 1, 2001, pp. 5–32.,https://doi.org/10.1023/a:1010933404324.

[7] Breiman, L. (1996). Bagging Predictiors. *Machine Learning*, 24(2). https://doi.org/10.1007/bf00058655

[8] Chen, J., Cui, T., Tang, J., & Song, Q. (2014). Remote sensing of diffuse attenuation coefficient using MODIS imagery of turbid coastal waters: A case study in Bohai Sea. *Remote Sensing of Environment*, 140, 78-93.

[9] Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu.* https://doi.org/10.1145/2939672.2939785

[10] Deepak R. Mishra and Sachidananda Mishra (2010) Plume and bloom: effect of the Mississippi River diversion on the water quality of Lake Pontchartrain, Geocarto International, 25:7, 555-568, DOI: 10.1080/10106041003763394

[11] Derot, J., Yajima, H., and Jacquet, S. (2020). Advances in forecasting harmful algal blooms using machine learning models: A case study with Planktothrix rubescens in Lake Geneva. *Harmful Algae, 99.* https://doi.org/10.1016/j.hal.2020.101906

[12] Ding, S., Chen, M., Gong, M., Fan, X., Qin, B., Xu, H., Gao, S. S., Jin, Z., Tsang, D. C. W., and Zhang, C. (2018). Internal phosphorus loading from sediments causes seasonal nitrogen limitation for harmful algal blooms.*Science of the Total Environment, 625.* https://doi.org/ 10.1016/j.scitotenv.2017.12.348

[13] Freund, Y., and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55(1).* https://doi.org/10.1006/jcss.1997.1504

[14] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29(5).* https://doi.org/10.1214/aos/1013203451

[15] Glibert, P. M., Anderson, D. M., Gentien, P., Granéli, E., and Sellner, K. G. (2005). The global, complex phenomena of harmful algal blooms. *Ocenaography,18(SPL.ISS.2).* https://doi.org/10.56

[16] Hill, P. R., Kumar, A., Temimi, M., and Bull, D. R. (2020). HABNet: Machine Learning, Remote Sensing-Based Detection of Harmful Algal Blooms. *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing, 13.* https://doi.org/ 10.1109/JSTARS.2020.3001445

[17] Izadi, M., Sultan, M., Kadiri, R. El, Ghannadi, A., and Abdelmohsen, K. (2021). A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom. *Remote Sensing* (Vol.13, Issue 19). https://doi.org/10.3390/rs13193863

[18] Kirk, J. T. O. (2011). *Light and Photosynthesis in Aquatic Ecosystems (3rd ed.).* Cambridge University Press.

[19] Kown, Y. S., Baek, S. H., Lim, Y. K., Pyo, J. C., Ligaray, M., Park, Y., and Cho, K. H (2018). Monitoring coastal chlorophyll-a concentrations in coastal areas using machine learning models. *Water (Switzerland), 10(8).* https://doi.org/10.3390/w10081020

[20] Li, X., Yu, J., Jia, Z., and Song, J. (2014). Harmful algal blooms prediction with machine learning models in Tolo Harbour. *Proceedings of 2014 International Conference on Smart Computing, SMARTCOMP 2014.* https://doi.org/10.1109/SMARTCOMP.2014.7043865

[21] Mahmudi, M., Serihollo, L. G., Herawati, E. Y., Lusiana, E. D., and Buwono, N. R. (2020). A count model approach on the occurrences of harmful algal blooms (HABs) in Ambon Bay. *Egyptian Journal of Aquatic Research, 46(4).* https://doi.org/10.1016/j.ejar.2020.08.002

[22] ] Masó, M., and Garcés, E. (2006). Harmful microalgae blooms (HAB); problematic and conditions that induce them. *Marine Pollution Bulletin, 53(10–12).*

https://doi.org/10.1016/j.marpolbul.2006.08.006

[23] ] Matthews, M. W., Bernard, S., and Robertson, L. (2012). An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters. *Remote Sensing of Environment, 124.* https://doi.org/10.1016/j.rse.2012.05.032

[24] McCorquodale, J. A., Roblin, R. J., Georgiou, I. Y., & Haralampides, K. A. (2009). Salinity, Nutrient, and Sediment Dynamics in the Pontchartrain Estuary. *Journal of Coastal Research, 10054.* https://doi.org/10.2112/si54-000.1

[25] Neves, R. A. F., Nascimento, S. M., & Santos, L. N. (2021). Harmful algal blooms and shellfish in the marine environment: an overview of the main molluscan responses, toxin dynamics, and risks for human health. . *In Environmental Science and Pollution Research* (Vol. 28, Issue 40). https://doi.org/10.1007/s11356-021-16256-5

[26] Robnik-Šikonja, M. (2004). Improving random forests. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 3201.* https://doi.org/10.1007/978-3-540-30115-8 34

[27] Roy, E. D., White, J. R., Smith, E. A., Bargu, S., and Li, C. (2013). Estuarine ecosystem response to three large-scale Mississippi River flood diversion events. *Science of the Total Environment, 458–460.* https://doi.org/10.1016/j.scitotenv.2013.04.046

[28] Sellner, K. G., Doucette, G. J., and Kirkpatrick, G. J. (2003). Harmful algal blooms: Causes, impacts and detection. *In Journal of Industrial Microbiology and Biotechnology* (Vol. 30, Issue 7). https://doi.org/10.1007/s10295-003-0074-9

[29] Smith, Emily Anne, "Cyanobacteria harmful algal blooms in South Louisiana estuaries: a synthesis of field research, management implications, and outreach" (2014). LSU Doctoral Dissertations. 283. pp. 16 – 55.

[30] Solomatine, D.P., & Shrestha, D.L. (2004). AdaBoost.RT: A boosting algorithm for regression problems. *IEEE International Conference on Neural Networks - Conference Proceedings,2.* https://doi.org/10.1109/ijcnn.2004.1380102

[31] ] Song, W., Dolan, J. M., Cline, D., & Xiong, G. (2015). Learning-based algal bloom event recognition for oceanographic decision support system using remote sensing data. *Remote Sensing, 7* (10).https://doi.org/10.3390/rs71013564

[32] *Spillway Operational Effects.* (n.d.). US Army Corp of Engineers New Orleans District Website. Retrieved February 3, 2022, from https://www.mvn.usace.army.mil/Missions/Mississippi-River-Flood-Control/Bonnet-Carre-Spillway-Overview/Spillway-Operation-Information/

[33] ] Steidinger, K. A., & Haddad, K. (1981). Biologic and Hydrographic Aspects of Red Tides. *BioScience, 31*(11). https://doi.org/10.2307/1308678

[34] Tang, W., Qin, J., Yang, K., Niu, X., Min, M., Liang, S. (2017). An efficient algorithm for calculating photosynthetically active radiation with MODIS products. Remote Sensing of Environment, 194, 146-154.

[35] Turner, A. D., Lewis, A. M., Bradley, K., & Maskrey, B. H. (2021). Marine invertebrate interactions with Harmful Algal Blooms – Implications for One Health. *Journal of Invertebrate Pathology, 186.* https://doi.org/10.1016/j.jip.2021.107555

[36] ] Xing, X. G., Zhao, D. Z., Liu, Y. G., Yang, J. H., Xiu, P., and Wang, L. (2007). An overview of remote sensing of chlorophyll fluorescence. In *Ocean Science Journal* (Vol. 42, Issue 1). https://doi.org/10.1007/BF03020910

[37] Yerrapothu, Bala Tripura Sundari, "Application of Machine Learning Techniques to Forecast Harmful Algal Blooms in Gulf of Mexico" (2021). Master's Theses. 809. https://aquila.usm.edu/masterstheses/809

[38] Yu, P., Gao, R., Zhang, D., and Liu, Z. P. (2021). Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators, 123.* https://doi.org/10.1016/

[39] Zheng, L., Wang, H., Liu, C., Zhang, S., Ding, A., Xie, E., Li, J., and Wang, S. (2021). Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models. *Journal of Environmental Management, 295.* https://doi.org/10.1016/j.jenvman.2021.113060

[40] Zingone, A., and Oksfeldt Enevoldsen, H. (2000). The diversity of harmful algal blooms: A challenge for science and management. *Ocean and Coastal Management, 43*(8–9). https://doi.org/10.1016/S0964-5691(00)00056-9

## Vita

Ian Mathew Smith, born in Baton Rouge, Louisiana, decided to go on several vacations with friends and enjoy life after obtaining their bachelor's degree from Louisiana State University in May 2019. Soon after, the pandemic of 2020 caused a hiring freeze that prevented him from finding a job in his field, and forced him return to serving tables at a local restaurant. These circumstances pushed him to consider bettering himself and expanding his knowledge in engineering and science. Having always considered going back to get a masters degree, Ian decided to enroll in a graduate program at Louisiana State University to obtain a degree in Civil Engineering concentrating in water resources. He plans to receive his Masters degree in August 2023, and after he will enter the work force as an full fledged engineer with a company that focuses on environmental and engineering tasks that better his community. Still having drive to learn, using his skills he obtained while in masters' degree program he plans to further his knowledge and obtain his professional engineering certification.