Random Forest Modeling and Long-term Water Quality Analysis for Algal Blooms in the Hudson-Raritan Estuary

Thesis Highlights:

- Analysis of a long-term dataset on water quality of the Hudson-Raritan Estuary, a severely degraded and understudied system adjacent to the highly developed New York City area.
- The development of a random forest model forecasts with reasonable accuracy the status of *Heterosigma akashiwo* one month in advance of the most recent sample collection. This model has potential implications in management.
- Evaluation of the random forest model gives insight into what variables are deemed important to the model, allowing hypotheses to be generated about biotic and abiotic relationships with *Heterosigma akashiwo* within the complex phytoplankton ecosystem.
- Long-term water quality monitoring data is used to search for trends in the condition of the Hudson-Raritan Estuary. Based on these trends, a general idea is provided of how the estuary is changing, likely as a result of anthropogenic activities.

Thesis by Evan Flint (‘23)

For the Academic Year (2022-2023)

Advisors: Dr. Megan Rothenberger

Committee: Dr. Trent Gaugler and Dr. Allison Lewis

Lafayette College Department of Environmental Science

Lafayette College Department of Mathematics

Table of Contents

Abstract

The Hudson Raritan Estuary (HRE), located south of New York City, has a well documented history of cultural eutrophication and associated harmful algal blooms (HABs). HABs in general can cause economic and ecological damage through development of hypoxic conditions and various toxins which in turn lead to fish kills and economic losses. To our knowledge, our lab has the longest continuous monitoring record of water quality and plankton in this system, representing 12 years of monthly sampling (April - November). Of the 13 species identified in the system as capable of causing harmful algal blooms, *Heterosigma akashiwo* was selected as the focus of the project. Prior studies in this system suggest *H. akashiwo* is the most frequent blooming species and regularly degrades the HRE by developing hypoxic conditions. The first objective of this paper is to use the random forest (RF) modeling technique to accurately forecast the occurrence of *H. akashiwo* blooms in the HRE to inform future management of the system. Dissection of the RF model will also provide insight into the best biotic and abiotic predictors of *H. akashiwo* blooms. The finalized RF model appears to predict correct classification of future HAB events moderately well (prediction error ~12%). Important predictors within the model include river discharge, precipitation, *Heterocapsa rotundata* abundance, *Chlamydomonas spp*. abundance, ferrous iron. The second objective was to summarize trends in water quality over the sampling period through nonparametric time series analysis using Seasonal Mann-Kendall and Theil-Sen's slope estimation. Results suggest several significant trends in water quality do exist over the course of our study including decreased dissolved oxygen and pH as well as increased ammonium and water clarity.

Biographical Sketch

I have always been very curious about my surroundings, searching to understand not only how things work but why they are the way they are. I love spending time outside, and so that is where my curiosity was most prominent, motivating my interest in the dynamics of ecosystems, evolutionary patterns, and ecological relationships between organisms. As I moved through my academic career, I realized that environmental science was the perfect interdisciplinary program where I could satisfy my curiosities and further apply what I learn to preserve the threatened natural world.

I also possessed a strong interest in mathematics, particularly in statistics as this was a way of understanding and interpreting the world that made a lot of sense to me. After deciding that I would officially double major between these two disciplines, I began looking for ways to combine them and discovered Dr. Rothenberger's lab. Dr. Rothenberger had been collecting data on the Hudson-Raritan Estuary for over ten years, culminating in a large dataset which was a perfect opportunity for me to apply my statistical background to an environmental problem.

I joined Dr. Rothenberger's lab my junior year and began working on my first big project. We decided to use a method referred to as ecological network analysis to investigate which of the many variables in the dataset appeared to be related, allowing us to build hypotheses and better understand the dynamics present in the ecosystem. I was lucky to help turn this into an official publication and got my first official authorship on a scientific paper.

My senior year I was intent on completing a thesis project and decided to continue working with this dataset and running statistics. I decided to take another approach to the data and look for a way to predict harmful algal blooms because they were a major threat to the

aquatic system we were studying along with many more around the world. As I read through papers on statistical analyses of harmful algal blooms I stumbled upon random forest modeling. As I discussed the concept with my mathematics professors, environmental professors, and some data scientists I realized this was exactly what I was looking for.

As I worked through the project, I learned even more about the challenges of constructing a proposal, presenting complicated ideas to a general audience, needing to have a full understanding of the topic (even outside of my specific disciplines), the peer review process, and more. This process also solidified my desire to continue my education with graduate school and pursue a research-based career in the future. After Lafayette College, I will be pursuing a PhD in Statistics at Oregon State University where I will focus on specifically applying statistical methods to environmental problems.

Introduction

1.1. Background

The major influence of phytoplankton in aquatic ecosystems have been well documented since the mid 1900s. Although they are primarily microscopic, phytoplankton are responsible for roughly half the global net primary productivity (Cloern et al., 2014; Field et al., 1998; Zhu et al., 2018). Phytoplankton also play a foundational role in aquatic food webs, feeding various species of fish and zooplankton grazers. Despite their positive effects on the global environment, dramatic shifts in local phytoplankton assemblages can have substantial negative ecosystem consequences. Events of rapid, localized phytoplankton growth are one such assemblage change which frequently have drastic consequences and are referred to as harmful algal blooms (HABs).

HAB effects generally fall into three overlapping categories: 1) ecological consequences on local flora and fauna, 2) human health and illnesses, and 3) economic. One of the most common effects of HABs is the development of hypoxic (or in severe cases anoxic) conditions. Three mechanisms lead to hypoxia: aerobic decomposition, increased turbidity, and increased respiration. Aerobic decomposition of phytoplankton is performed by bacteria upon death of phytoplankton cells. As large volumes of phytoplankton die off, bacteria siphon dissolved oxygen out of the water to break down the cells (Suthers & Rissik, 2009; Yang et al., 2019). Increased turbidity refers to a decrease in water clarity, in this case resulting from higher phytoplankton density within the water column. This inhibits sunlight penetration to submersed aquatic flora, and thereby lowers photosynthetic rates of species including seagrasses which act as a major source of dissolved oxygen (Suthers & Rissik, 2009). Not only can this lower photosynthetic rates in existing plants, but cultural eutrophication and algal blooms have been

linked to seagrass disappearances in coastal waters (Burkholder et al., 2007). Lastly, increased respiration rates result from the increase in phytoplankton abundance. As with most other photosynthetic organisms, the absence of sunlight at night prevents photosynthesis from occurring, and thus cellular respiration is performed. This process demands oxygen which is derived from the surrounding water.

Hypoxic conditions generated by these mechanisms lead to the differential suffocation of finfish and shellfish. Observation of the HRE native species striped bass (*Morone saxatilis*) and bay anchovy (*Anchoa mitchilli*) within the Chesapeake Bay ecosystem exhibit relatively low tolerance to decreases in DO, with LC_{50} equal to about 2.5 mg/L and 2.8 mg/L respectively (Berg & Levinton, 1985; Breitburg et al., 2001). Shifts in DO can thus lead to shifts in fish community structure with cascading implications for non-fish species at other trophic levels.

Another well documented result of HABs is the release of toxins from certain species. Evolution appears to have selected for toxins for both defensive and offensive functions. Defensively, secreted chemicals can suppress competitors and defend against predation from grazers (Driscoll et al., 2016; Legrand et al., 2003). Offensively, these chemicals can act as venom to immobilize prey (Sheng et al., 2010). That said, the nature of toxin production as an evolutionary response to other organisms makes these physiologies very challenging to study using both in-situ and ex-situ methods as they are inherently based on interactions within complex aquatic communities. HAB events composed of toxin-releasing taxa can beget accumulation of toxins in the ecosystem. Some of the most common incidents of phytoplankton-produced toxins along the eastern coast of North America are amnesic shellfish toxin (AST), ciguatera shellfish toxin (CST), diarrhetic shellfish toxin (DST), neurotoxic

shellfish toxin (NST), and paralytic shellfish toxin (PST) (Hallegraeff et al., 2021). These toxins are known to accumulate in finfish and shellfish, causing mass mortality events when concentrations exceed lethal thresholds (Jin & Hoagland, 2008; M. B. Rothenberger et al., 2014). Consumption of fish originating from fisheries or aquaculture operations impacted by these toxins can lead to the ingestion of toxins by the general public. Symptoms range from relatively benign cases of vomiting and abdominal discomfort to more severe organ failure and respiratory complications, even leading to death in extreme cases (Pettersson & Pozdni͡akov, 2013; Sarkar, 2018; Suthers & Rissik, 2009).

Physiological impact of phytoplankton on fish, particularly finfish, are less often discussed but still important. Several species of plankton possess spine-like anatomical structures. When these species get lodged in the gills of finfish, they can cause discomfort, irritation, and abrasions of the gills. Resulting inflammation can increase susceptibility to infection (Pettersson & Pozdni͡akov, 2013; Suthers & Rissik, 2009). .

The aforementioned hypoxic/anoxic conditions, toxin accumulation, and physical damage have pronounced influences on local wildlife. Fish kills are commonplace results of both situations, and mortality events at one level of the trophic web generally cascade to other sectors, resulting in widespread degradation of ecosystem services. The collapse of fisheries and tourism are two ecosystem services with economic importance for many coastal areas threatened by HABs. In 2005, an algal bloom off the New England coast resulted in losses from the softshell clam, mussel, quahog fisheries that cost the states of Massachusetts and Maine roughly $15.7 million and $2.5 million respectively (Jin & Hoagland, 2008).

1.2. HAB Causes and Dynamics

Public pressure to resolve the HAB issues in coastal areas has led to intensified monitoring and research efforts on the topic. Improvements in algal bloom monitoring technology and more long-term studies are resulting in accumulations of volumes of data in databases such as the Harmful Algal Event Database (HAEDAT). These large datasets provide the ideal opportunity to investigate HAB dynamics with the use of advanced statistical techniques. Certain statistical techniques, including the random forest model used in this study, can provide forecasting information, potentially informing ecosystem managers of HAB events and creating an opportunity for proactive style management instead of the typical reactive style. Proactive management can aid in the mitigation or prevention of the consequences of HABs.

A majority of the research has focused on abiotic variables such as nutrient loads and water quality parameters, and weather/climatic influence. Chemicals like P, N, and Si, are known to be limiting and required in abundance for phytoplankton growth, whereas other nutrients including Cd, Co, Ni and Zn are also known to be absorbed by phytoplankton but in smaller amounts (Hecky & Kilham, 1988; Paulsson & Widerlund, 2021; Suthers & Rissik, 2009).

Nitrogen, often as nitrate and ammonium, and phosphorus, often as phosphate, are particularly important because in addition to being limiting nutrients, they are components of common pollutants including sewage and fertilizer runoff. This process of anthropogenic nutrient addition resulting in phytoplankton growth stimulation and ecosystem degradation, or cultural eutrophication, is one of the most general ways nutrients in a system can influence phytoplankton assemblages. Other influences can be more species specific, such as decreased

Si:N ratios tend to shift assemblages away from diatom dominance, and pulses of iron into a system can result in dinoflagellate dominance (M. B. Rothenberger & Calomeni, 2016).

Weather and climate patterns also play an important role in phytoplankton patterns, as previously mentioned in regard to projected HAB patterns with the progression of climate change. Several influential weather patterns relate to nutrient availability, such as precipitation events often correlating positively to phytoplankton growth due to increased runoff transporting nutrient rich fertilizers and sewage into aquatic systems. River discharge is similarly correlated with phytoplankton growth and reproduction for related reasons. The relationship between HABs and weather/climate and nutrient availability is clearly crucial to understanding the dynamics of phytoplankton and is well understood and thoroughly researched in algal literature.

Conversely, the information on biotic interactions between species of plankton, fish, and bacteria is only superficially understood. In a 1961 publication entitled *Paradox of the Plankton*, the author motivates the necessity to understand these biotic relationships and puts forth that HABs cannot be fully understood without them. With a limited amount of nutrients and a vast collection of phytoplankton species competing for them, principles of competitive exclusion would predict a far less diverse plankton community than actually observed (Hutchinson, 1961). Hutchinson continues by proposing the reason such high levels of diversity are achieved relates to the symbiotic relationships present between species, and thus understanding these relationships is extremely important in understanding overall HAB dynamics.

The gap in knowledge on biotic relationships is likely a consequence, in part, of the complexity of phytoplankton communities which can be composed of a sizable set of species. Previously mentioned advancements in monitoring technologies, techniques, and statistical

analysis methods have only recently provided an opportunity to truly understand these communities. Furthermore, relationships among species are not easily generalized between systems, and so these relationships must be studied on a site-by-site basis (Hallegraeff et al., 2021).

Despite these challenges, we have some generalizable knowledge on relationships at the broader taxonomic levels. For instance, zooplankton are known to graze on phytoplankton, a dynamic which can potentially be utilized for HAB management (Griffin et al., 2001; Reynolds, 2006; Schoenberg & Carlson, 1984). Bacterial relationships to phytoplankton are less well understood but it is acknowledged that they play an important role in the microbial loop of aquatic systems (Suthers & Rissik, 2009). Studies suggest important symbiotic relationships exist and are postulated as being either growth promoting (obligate or facultative) or growth inhibiting (Seymour et al., 2017).

1.3. Advancements in HAB monitoring

Forecasting ability for HAB activity is informed by the wealth of knowledge already available on the topic, particularly pertaining to what variables drive or inhibit bloom formation of local species. Satellite imagery and more precise equipment are two major improvements in the field, allowing scientists to compile vast quantities of data for analysis. The development of global databases like HAEDAT is also a major improvement, allowing scientists to conveniently access data.

Consequently, scientists studying aquatic ecosystems are able to detect fine alterations in HAB patterns, and some studies suggest HABs have been increasing in both intensity and

severity (Anderson et al., 2021). The legitimacy of this proposition is uncertain, as Hallegraeff et al. (2021) point out, since more frequent and reliable monitoring methods could confound these trends. As monitoring methods improve in accuracy, blooms may be detected which may have been missed using old methods, and so although blooms would occur with the same frequency monitoring would detect an increase in bloom activity.

Regardless, studies attempting to forecast shifts in HAB patterns reveal further increases in HAB frequency and intensity are likely with the progression of climate change. Elevated oceanic carbon dioxide concentrations, rising sea temperatures, and more frequent upwelling events are all favorable conditions for certain HAB-forming species (Hallegraeff et al., 2021; Sarkar, 2018). This further underscores the importance of understanding phytoplankton community dynamics and developing methods to predict HAB occurrence. Accurate prediction can allow ecosystem managers to proactively control HAB events and mitigate the potential harms of the forecasted bloom.

1.4. The Hudson-Raritan Estuary

Broad understanding of phytoplankton and HAB dynamics is crucial foundational knowledge, but differences in climate, hydrology, and ecology between ecosystems necessitates a site-by-site analysis for understanding a particular system (Hallegraeff et al., 2021). This study focuses on the Hudson-Raritan Estuary (HRE), a brackish water system located between the states of New York and New Jersey, directly south of New York City. Before development, the HRE was an important recreational fishery, supporting local finfish, shellfish, and waterfowl (Kane & Kerlinger, 1994). In the past decade and a half, declines in the fisheries have been

observed, possibly a combined effect of HABs, pollution, and overfishing. Collapse of the eastern oyster (*Crassostrea virginica*) fishery is a clear example of this as the species is currently considered ecologically extinct in the HRE (Jeffries, 1962). Other evidence of ecosystem decline is the fall of commercial catch of American shad as well as the deterioration of the sturgeon fishery by the early 1900s (Berg & Levinton, 1985).

Proximity to the city has led to intense coastal development and changes in hydrology which, in conjunction with pollution from river outflow, created eutrophic conditions since at least the 1960s (Jeffries, 1962; Rothenberger et al., 2018). The 1962 study by Jeffries was a baseline for comparison for the data collected by our lab. Findings suggest 2014 SRP concentrations were up to 20 times higher than in 1962 at some sample locations and nitrate concentrations are up to 50 times higher, causing the HRE to continue exhibiting cultural eutrophication symptoms like frequent algal blooms and seasonally low dissolved oxygen levels (Jeffries, 1962; Rothenberger et al., 2014). High concentrations of other pollutants including PCBs, heavy metals, and pesticides have also been detected in the system (Breteler, 1984).

The HRE has been studied by our lab since 2010, representing the longest known modern study of water quality and plankton relationships in the ecosystem. The most recent publication is an analysis of all collected data using ecological network analysis to reveal correlations between species abundances, weather factors, and nutrient loads with a primary goal of filling the knowledge gap on biotic relationships in plankton assemblages (Rothenberger et al., 2023). Ecological network analysis was able to uncover correlations between variables existing at a given temporal sample. These correlations have potential to generate hypotheses, specifically about the existence and importance of symbiotic relationships between species. We found about

95% of the revealed correlations were biotic, supporting the hypothesis that biotic drivers are likely very important in this system (Rothenberger et al., 2023). This prior study motivates the current one, where our new analysis objectives focus on digging deeper into causal relationships of HABs of a particular species. The highly interconnected nature of phytoplankton communities revealed within the HRE supports the need for modern statistical analysis methods to investigate relationships and HABs within the system. Our first objective of constructing a predictive model directly addresses this, and the second objective of overall water quality analysis plays a supplementary role in understanding changes in the estuary which may in turn influence phytoplankton communities.

Several phytoplankton taxa have been recorded as forming HABs in the HRE including *Ceratium tripos, Prorocentrum micans, Heterocapsa rotundata,* and *Heterosigma akashiwo* (Rothenberger et al., 2023). Over the course of our sampling, the species *H. akashiwo* has bloomed most frequently and for this reason was chosen as the focus of this study (Rothenberger et al., 2023). *H akashiwo* is a raphidophyte known to release brevetoxin in other ecosystems with effects on fish including paralysis and mortality; however, to our knowledge, brevetoxin has not been recorded in the HRE. (Graham & Wilcox, 2000; Khan et al., 1997). Ingestion of brevetoxin by humans is known to cause reversed hot and cold sensations, vertigo, vomiting, tingling sensations, and abdominal pain (*CDC | Case Definition*, 2019). The most common vector of exposure is through consumption of contaminated shellfish with brevetoxin bioaccumulation. Such impacts of *H. akashiwo* blooms necessitate a better understanding of bloom events which can be accomplished by more advanced statistical methods.

Methods

2.1. Sampling

Field sampling in the HRE occurred by boat between 2010 and 2022 with samples collected at monthly intervals between April and November (weather prohibited sample collection in other months) between 1000 and 1200 hours. Six sites were chosen for sampling (**Figure 1a**) in the estuary. Selection of these sites was to compare modern results to those of Jeffries from half a century earlier. Reasoning for the selection of these sites in the 1960s was motivated by circulation patterns pervasive in the estuary which modern studies confirm persist (**Figure 1b**) (Jeffries, 1962).

At each sample location, water samples were collected at three depths referred to as surface (0.5 m depth), middle (1.5 m depth), and lower (3.0 m depth) with a thoroughly rinsed 2.2 L Van Dorn water sampler. Water was then transferred into acid cleaned bottles for analysis with SRP being first filtered through a sterile 0.45 μm syringe filter. SRP, nitrate, and ammonium were collected at all three depths while ferrous Fe and silica were only collected at the surface. Water temperature, pH, dissolved oxygen concentration, and conductivity were measured at each depth (surface, middle, and lower) with a YSI 6820 V2 multiparameter meter calibrated before each sample session. Water clarity was estimated using a Secchi disk, a black and white disk which is lowered into the water column to measure turbidity. The depth at which the disk disappears is the measurement recorded. A 12-1 Schindler-Patalas trap was used on the surface to collect a zooplankton sample. Phytoplankton was collected from the surface Van Dorn sample and preserved with acidic Lugol's solution in amber bottles to prevent shock which can complicate visual identification. All plankton were held at 4°C until analysis.

Nutrient analysis in the laboratory was performed according to standard methods (Rice et al., 2012). A HACH DR/2500 spectrophotometer was used for analysis, calibrating with standards before each session and blanks between samples. Nitrate and ferrous Fe were performed within 24 hours after collection and SRP, silica, and ammonium were performed within 48 hours. If sample analysis was not possible within the timeframe, samples were frozen and thawed when able to be analyzed. Plankton analysis was performed by cell counts under microscopy identified to the lowest taxonomic levels and recorded as density. Weather data were acquired from New Jersey Weather and Climate (weekly precipitation), Rutgers Office of the New Jersey State Climatologist (monthly precipitation), and the USGS (river discharge). All directly observed variables (nutrients, plankton abundances, weather, etc) were included and variables observed at multiple depths were included as separate variables (i.e. $[NO_3^-]$ at surface, $[NO_3^-]$ at middle, and $[NO_3^-]$ at lower). All data were recorded in a collection of spreadsheets for future analysis.

2.2. Data Organization and Software

Data analysis was performed in RStudio with the packages *ranger* for the analysis described later and *missForest* for imputing missing values and all aforementioned variables were included in statistical analysis. The dataset was then structured to achieve one of the primary objectives of the study, predicting algal blooms of *H. akashiwo*. To accomplish this, data were reorganized in a way that associated one month's *H. akashiwo* abundance with previous month's variables.

The lag period was chosen to be 2 months to balance our chances of capturing important variables with the loss of samples early in the year. Yajima and Derot (2018) ran random forest analysis on chlorophyll-a concentrations as a proxy for algal abundance and tested lag times of 1, 2, and 3 months. They determined a 2 month lag to perform best, however results were similar between the groups. Because two months of prior data are needed to build the model, observations of *H. akashiwo* within the first 2 months of our sample season cannot be predicted (e.g. April could not be predicted because we did not have data on February and March to lag). From a management perspective, an effective model requiring minimal data collection is also optimal.

Missing data were present within the dataframe for a variety of reasons (suspended sampling, equipment malfunction, etc.). For the random forest model, imputation was performed with the package *missForest* in RStudio for any observations with minimal missing data. For the time series analysis of overall trends, missing data were ignored. Any full month without data collection was fully ignored.

2.3. Random Forest Modeling

The primary method of statistical analysis for this study was construction of a classification random forest (RF) model. RFs are an ensemble method, combining many individual decision trees built from a training dataset to predict an unknown variable (Breiman, 2001). Decision trees segment p variables in the predictor space and group response values based on that split. The ideal split is chosen at each node, where ideal splits are those which group observations into most similar groups, and splitting continues until each group has some small

number of observations filtered into it, called a terminal node. An overfitting issue, pervasive in decision trees, can be addressed by introducing bagging, referring to the combination of bootstrapping the training data to construct multiple trees based on a random subset of $p' = \frac{2}{3}p$ predictors, and aggregating, averaging the responses of B individual trees to calculate the overall response of the forest given by

$$\widehat{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} f^b(x)$$

where $\widehat{rf}(x)$ is the response of the forest and $f^b(x)$ is the response of a single tree. Further, at each split $m \approx \sqrt{p'}$ variables are randomly selected from the total set and evaluated for the best possible split. The following spit will then evaluate another m variables randomly chosen and this proceeds until terminal nodes of the specified size are achieved. This process serves to decorrelate trees from one another, addressing the overfitting issue observed in single decision trees.

This paper uses a classification RF model which yields a categorical response variable (opposed to a regression RF model which yields a continuous response). Although a regression RF model was initially tried, the classification method was chosen due to the nature of the response variable which, although recorded as a continuous variable in cells per mL, exhibited a highly skewed distribution. Other studies on similar datasets have also chosen to use the classification method due to its outperformance of the regression style (Derot et al., 2020; Harley et al., 2020). Use of a regression style forest generally led to very poor predictive ability with extremely high error rates. Classification of *H. akashiwo* abundance was done by quartile ranges

where values below the 90th percentile were considered "low" abundance, and those above were considered "high". For our dataset, this amounts to the cutoff level being 852 cells/mL. **Figure 2** shows a time series plot of *H. akashiwo* over the period of study which was used to decide the cutoff ranges.

Evaluation of the model was investigated by out-of-bag error (OOB) and a confusion matrix. OOB error works by testing each tree with data not used in its creation, made possible by the bootstrapping process which subsets the data. Thus some observations are unused in the construction of a given tree and we can calculate the error for that tree as whether or not it makes an accurate prediction. Because this is a classification forest, the Gini index (G) was used as a measure of error within the trees. This can be thought of as a measure of node purity. Recall that RFs split up observations by categorizing them into like groups based on trends in predictor variables. However, that terminal node usually groups together some number of observations greater than 1, and so the most common response of the observations in that group is declared the response of that terminal node. This method is also referred to as "majority vote". When the responses of the observations in the terminal node are more homogeneous, the node is deemed more pure. Then, if we let g_k be the proportion of observations from the dominant class in the k^{th} terminal node,

$$G = \sum_{k=1}^{K} g_k(1 - g_k)$$

This equation implies as terminal nodes become more homogeneous, g_k is closer to 1. Thus G approaches 0 representing better forest classification. m^th^

The confusion matrix displays the true category of an observation and the category predicted by the forest for that observation. Similar to the OOB error, this method evaluates by running each observation only through trees created without that observation. This gives an idea of how well the model is classifying observations of specific types as true positive (TP), true negatives (TN), false positives (FP), and false negatives (FN). The two metrics calculated from this table are accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

and precision

$$\text{Precision} = \frac{TP}{TP+FP}$$

Lastly, variable importance was calculated as a means for interpreting the RF. Variable importance within a tree was ranked by reduction of the Gini index at each split. For a given predictor, the Gini index is calculated both before and after every split in every tree which uses that variable. The reduction in split is then averaged across trees, and variables which show a greater average reduction of the Gini index are deemed to be more important, as they do a better job at homogenizing the terminal nodes of the RF.

Although variable importance gives a general idea of what variables are more crucial in creating accurate predictions within the forest, the nature of the relationships between variables is obscured. It cannot be easily determined what role each variable plays within each tree where it is used as a predictor. The number of trees within the forest also obscures these relationships, as they may exhibit slight variations based on what variance is previously accounted for by other

variables within a single tree. This is a major drawback of this method, but the obscurity is the cost for improved accuracy of the model with real-world data.

2.4. Time Series Analysis

Summary statistics were performed to investigate overarching trends in several important water quality parameters of the HRE. Overall trends were analyzed using data from the 2010-2020 time period due to frequent sampling disruptions from 2020-2022 (a consequence of COVID-19). Analysis of the long-term trends throughout the 10 year study period were performed as well as analysis for fluctuation of parameters by season.

Long-term trends were analyzed using Seasonal Mann-Kendall from the *rkt* package in RStudio. This is a nonparametric approach to identify monotonic trends in the data over the time series for which the null hypothesis is randomness while the alternative is the existence of a monotonic trend. Selection of this test was based on the nonparametric nature of the test, as well as its ability to cope with missing values and frequent use in environmental time series literature (Hirsch et al., 1982). Seasonal Mann-Kendall identifies a monotonic trend throughout a whole time series by first comparing n annually recurring observations, $\{x_1, x_2, \cdots, x_n\}$, for each season individually (i.e. the trend in April is calculated then separately the trend in May is calculated and so on for each season of the time series). For each season, we can calculate the trend (S) as

$$S = \sum_{i=1}^{n+1} \sum_{j=1}^{n} sgn(x_j - x_i)$$

where

$$sgn(\theta) = \begin{cases} 1 \text{ for } \theta > 0 \\ 0 \text{ for } \theta = 0 \\ -1 \text{ for } \theta < 0 \end{cases}$$

Effectively, this test gives a +1 ranking when one observation is higher than the previous, a -1 ranking when a season is lower than the previous, and 0 when equal, such that the overall test statistic gives an idea of how strong the trend in data for a particular season is over time.

Once the trends for individual seasons are calculated, they can be aggregated across m seasons in the time series by

$$S' = \sum_{i=1}^{m} S_i$$

The statistics S' now gives an evaluation of the trend throughout the entire time series. This in turn allows for the calculation of the test statistic Kendall's Tau ($-1 \leq \tau \leq 1$) by

$$\tau = \frac{S'}{n(n-1)/2}$$

This effectively serves to evaluate trends over time and remove seasonality influence. However, if there are more complicated time series trends (i.e. an increase in some months coupled with a decrease in others), this method may miss those trends.

Theil-Sen's slope was also estimated with the *rkt* RStudio package to quantify the estimated change over the course of the study. This method calculates the slope for each season by ordinary least squares regression and takes the median. Similarly, the intercept for the Theil-Sen's slope is calculated as the median of the OLS regression intercepts. Use of this method is commonly used in conjunction with general Mann-Kendall methods due to its ability

to handle nonparametric data and resistance to outliers (Meals et al., 2011). Both Seasonal Mann-Kendall and Theil-Sen's slope were calculated for dissolved oxygen, pH, salinity, ammonium, nitrate, SRP, N:P ratio, Si:N ratio, and Secchi depth. Data was aggregated between sites and separated by depth, following procedure from (Burkholder et al., 2006).

Results

3.1. Random Forest Tuning

The different model parameters were evaluated to determine the most effective model. It was found 500 trees were sufficient as error reduction leveled off around 300 (**Figure 3**) trees for all parameter combinations. Using a 500 tree forest, iterations were done to minimize OOB error by evaluating different combinations of three main variables: size of the bootstrapped sample as a percent of the total observations (p_s), variables tried at each split (p_t), and minimum node size (p_n). Since general guidelines of values are provided for each of these, only slight variations on each were evaluated for reduction of OOB error (**Table 1**).

In addition to the standard parameters, a weighting factor was considered. Classification of *H. akashiwo* abundance into the two categories resulted in a distribution of 44 observations in the "high" abundance classification and the remaining 235 observations as "low". Early iterations of the model found decent predictive ability (around 15% error rate), however evaluation of the confusion matrix found extremely low accuracy and precision for high density events. This was likely due to systematic general classification of observations as low, yielding many correct responses for low observations but most of the few high observations were misclassified as low.

To address this, various weightings were tested (**Table 2**) to determine the optimal weight to place on correctly predicting higher algal density events. Effectively, adding a weight parameter tells the model that it is more important to correctly identify observations of "high" algal density. Although this typically improves the precision of the RF model for predicting high

density events, this may come at the cost of a loss in precision for low density events or overall prediction error. Costs and benefits of the weighting parameter can be seen in **Figure 4**. Prediction error and accuracy of the RF model appeared to be optimal in the 3-7 weighting range. As expected, precision of "high" predictions appeared to continually increase and precision of "low" predictions continually decreased, although improvement in "high" precision was far greater than the loss of precision for "low" predictions.

A weight of 6 was ultimately decided for the "high" classification, as "high" classification precision appeared to level out at this point and both accuracy and error were optimal. Further, the deterioration in "low" precision was able to be mitigated more than if a weight of 7-10 was used. Using this weight, the final RF model was developed and optimal parameters were determined from **Table 1** to be $p_t = 21$, $p_n = 9$, and $p_s = 0.65$ by permutation of all combinations and choosing that with the lowest OOB prediction error (11.83%). This RF model yielded the confusion matrix presented in **Table 3**. Accuracy was calculated as 88.17%. Precision for "high" and "low" counts were calculated as 63.64% and 92.77% respectively.

3.2. Variable Importance

The variable importance plot for the final RF model is shown in **Figure 5**, where it should be noted that only the top 25 of the +350 variables tested were included in the plot. Important results from **Figure 5** are the distribution of variable type (biotic, chemical, or weather parameter) and the actual variables present.

Weather variables of both lag times appeared to be highly important in predicting a *H. akashiwo* HAB event. Precipitation, measured as total monthly and weekly rainfall, and mean discharge of the Raritan River were all present in the top 25 important variables. Water geochemical parameters are also common throughout the plot. Dissolved oxygen for several depths from both lag times appears to play an important predictive role in the model. The only nutrient to appear in the figure is ferrous iron for which the surface measurements at lag one and the middle measurements at lag two were deemed important. Nitrate, ammonium, and SRP did not appear in the top 25 important variables at any depth or lag time. Lastly, several biotic variables appear to be important in the RF model. The biotic variables which appeared were *Heterocapsa rotundata* abundance at lag two, *Chlamydomonas spp*. abundance for both lag one and two, cyanobacteria abundance at lag one, *Leptocylindrus minimus* at lag one, and *Copepod nauplii* at lag two.

The overall pattern of the plot also tells us that most variables within the top 25 are similar in importance. Besides the top 5, the change in importance between any two adjacent variables is minimal, and considering error it would be difficult to legitimately assert that one is definitely more important than the next. Within the top 5 however, there is a clear structure to the importance of these variables to the RF model's predictive ability.

3.3. Long-term Trends

Performing the Seasonal Mann-Kendall test and calculation of the Theil-Sen's slope are summarized in **Table 4**. Four out of the nine variables showed significance (determined by

$\alpha = 0.05$) for at least one depth with non-zero Theil-Sen's slope estimations. Of the variables that showed significance for at least one depth, all three depths were found significant.

The time series plots for the four variables of significance are illustrated along with their Theil-Sen's derived trendlines in **Figure 6**. Dissolved oxygen surface and middle depths showed the strongest significance, as well as the Theil-Sen's slope of the greatest magnitude. pH showed slight significance at all three depths, also with negative slopes of comparable values between -0.05 and -0.07. Ammonium showed strong significance at surface and middle depths with moderate significance at the deep sample depth. A positive slope was found to be between 10 and 20 µg/L change per year for all depths. Finally, Secchi depth showed strong significance with a slight positive slope of 0.12 m per year.

As seen in **Figure 6.c**, apparent outliers for ammonium concentration occur between 2019 and 2020. Analysis with Seasonal Mann-Kendall and Theil-Sen's slope were both performed with the outliers removed (2010-2019) for the ammonium data, however it should be clarified that official analysis of ammonium data includes the outliers because we could not find any reason to officially exclude them. However, the outliers are so drastic that it is worth analysis of trends without them. For the surface, middle, and deep depths respectively a rerun found Kendall's Taus = 0.369, 0.336, and 0.170, p-values = 0.0003, 0.0009, and 0.0972, and Theil-Sen's Slopes = 13.33, 13.08, and 7.13. Kendall's Taus are slightly lower without the outliers, but still suggest a positive trend in ammonium concentration. Slopes without outliers are within 5 units below the slopes with outliers included, suggesting the positive trend in ammonium may be of a slightly smaller magnitude. Interestingly, while p-values for surface and

middle depths continue to show very strong significance, the deep depth fails to show significance at $\alpha = 0.05$.

Time series analysis was also run for *H. akashiwo* abundance over the course of our study (**Figure 2**). Two obvious periods of high abundance occur in the summers of 2012 and 2013, and a smaller peak appears in the summer of 2017. A box plot showing season-to-season differences in *H. akashiwo* abundance (**Figure 7.**) emphasizes the seasonality of species abundance. Algal density of *H. akashiwo* systematically begins low in the early spring, and begins to increase into the warmer months. After peaking around June or July, abundance begins to decrease again to early spring levels before sampling is suspended for the winter months. The summer months also display more variability compared to the months early and late in the sampling season.

Discussion

4.1. Implementation of the Model

The RF model developed in this paper, while far from perfect, does a good job predicting the observations with which it is unfamiliar, exemplified by the relatively low out-of-bag error. Other publications which attempted to establish RF models for HABs with various methodologies have had mixed success. Derot et al. (2020) performed a classification RF model and had an out-of-bag classification error (a measure of incorrect prediction frequency) of about 2% while the classification RF constructed by Harley et al. (2020) to predict paralytic shellfish toxin only showed an error rate of closer to 20%.

The weighting factor for "high" abundance observations (above the 90th percentile) evaluated in addition to the permutation of parameters allowed the model to be tuned for optimal performance. The expectation was for prediction error and accuracy to both exhibit deterioration as weighting for "high" classification was increased. It was reasoned that under the unweighted condition, the forest would perform optimally overall creating the forest with the lowest error. The addition of the weight parameter would then improve precision for "high" classification at the expense of "low" classification error which would eclipse the improvements from "high" precision. Although this was observed, the loss in precision for "low" classification was initially so minimal compared to the great improvement in precision for "high" classification, that error and accuracy showed initial improvements before deteriorating.

We expect that this can be explained by the low fraction of "high" observations in the dataset (44 of 279). Since "high" observations are so infrequent, they are often missed in the

random sampling involved in the bootstrapping step of the forest. The weighting factor overcomes this challenge such that the final model is able to predict both classes well. We recommend that future RF models established for HAB datasets incorporate this parameter as this leads to an overall better model.

Based on the relatively low error rate, this model has potential for application in the HRE ecosystem. Although monitoring of the HRE has concluded for our lab, several governmental and nonprofit organizations manage the HRE and could potentially utilize this model as a tool for prediction of *H. akashiwo* HAB activity. The US Army Corps of Engineers and the New York/New Jersey Port Authority outlined a restoration plan in 2020, one of the goals being to restore the eastern oyster populations in the estuary with roughly $3 million set aside for continued monitoring of the ecosystem (*Hudson-Raritan Estuary Ecosystem Restoration Feasibility Study*, 2020). The depletion of these oyster beds is a major concern, however reintroduction of the species will likely fail so long as *H. akashiwo* blooms fail to be controlled. The eastern oyster tends to exhibit elevated lysosomal destabilization rates, a metric of cellular damage, in response to *H. akashiwo* both in field and laboratory settings (Keppler et al., 2005). Thus, the ability to predict and manage blooms is both feasible and important, as money is being allotted for HRE monitoring and restoration projects are dependent on HAB control.

Additionally, nonprofits like the Harbor and Estuary Program, part of the Hudson River Foundation, are working to restore the overall ecosystem and improve the recreational and economic value (*Hudson River Foundation*, 2022). Implementation of this predictive model can help inform tourists when visitation is likely to be disrupted by an algal bloom, hopefully increasing tourism as recreational activities can be better scheduled.

Another important implication of the model could be for human protection from shellfish contamination. Although our monitoring has not observed brevetoxin within the HRE, we have not been testing for it and so it potentially is present. By implementing this model, at time when a bloom of *H. akashiwo* is soon predicted, inspection of fishery yield can be enhanced to assure contamination does not occur,

Lasty, it should be restated that this model can be used exclusively to predict HABs of *H. akashiwo*. In theory, similar models could be developed for other species known to form HABs on this system, and it may be beneficial to develop models for other species identified on the HAEDAT. Additionally, although HABs can be monospecific, they can also be composed of several species. HABs of this nature will not be effectively predicted by the model developed in this paper and better identified using chlorophyll-a measurements which are not a component of our dataset.

4.2. Predictor Relationships

The variable importance plot provides insight into which variables are best used to predict the abundance of *H. akashiwo* blooms in the near future. The six weather variables included in the dataset (prior month total precipitation, prior week mean precipitation, and prior week mean discharge) are highly important for predicting bloom activity. Precipitation will generally lead to an increase in discharge, and discharge is recognized as a highly influential factor in estuarine systems because of the transport of nutrients from upstream (Zhu et al., 2018). Nutrients like nitrogen, phosphorus, and ammonium can be transported from wastewater and

fertilizer runoff from upstream areas and deposited within the estuary, stimulating the growth of specific species.

Interestingly, nitrogen and phosphorous variables did not appear in the top 25 important variables in **Figure 5.** These nutrients are widely recognized as limiting plankton growth and stimulate algal blooms when deposited in large quantities. It is possible that our lag times were ineffective at capturing the influence of these variables on phytoplankton populations. Although it depends on the sample site, Rantajärvi et al. (1998) examined temporal sampling distribution and concluded that intervals as short as weekly would be needed to accurately capture phytoplankton dynamics in their study system. Because of the distance of our study site and the lack of a remote monitoring setup, this was not feasible for this study and monthly data was collected instead. The monthly sampling style and absence of data over wintering months made a two month lag optimal for our model. Also, it is possible that the influences of discharge and precipitation already accounted for variation which would have been explained by these nutrients. As mentioned, precipitation and discharge generally stimulate algal blooms by creating an influx of nutrients. However, correlations between weather variables (discharge and precipitation) and nutrients (ammonium, nitrate, iron, SRP, and silicon) do not seem to be supported by our data and so this explanation is not supported.

Another possibility is that these nutrients are not limiting for *H. akashiwo* and thus would play a relatively insignificant role in predicting *H. akashiwo* blooms. In eutrophic systems like the HRE, nitrogen and phosphorus are generally very available, and this may lead to other nutrients becoming limiting. The roles of nitrogen in amino acids, nucleotides, and chlorophyll as well as the role of phosphorus in ATP, DNA, and phospholipids are extensively documented

and recognized (Graham & Wilcox, 2000). The concept of iron as being a limiting nutrient in eutrophic systems has received comparatively little attention. Iron plays a critical role in photosynthesis and the construction of enzymes which facilitate the uptake of nitrate (Graham & Wilcox, 2000; Shaked & Lis, 2012). Evidence of iron as a limiting nutrient has also been observed in many field studies which use this to explain a phenomena of high nitrogen and phosphorus loads but relatively low phytoplankton loads (Behrenfeld et al., 1996; Hutchins et al., 1998). Yamochi et al. (1983) identify pulses of iron into the Osaka Bay of Japan as a driver of *H. akashiwo* bloom formation. Our findings of ferrous iron as one of the most important predictors in the RF model corroborate these results and emphasize the importance of considering ferrous iron as an important algal bloom driver in eutrophic ecosystems.

Also important to note is the dramatic spikes in *H. akashiwo* abundance in 2012 and 2013 were immediately preceded by intense hurricane events. Hurricane Irene (August 2011) and Hurricane Sandy (October 2012) both led to dramatic increases in discharge and precipitation. The peak of the 2011 bloom was more than six times the base level for that season, and the spike in 2012 was more than double the base level. Rothenberger et al. (2018) studied the influence of Hurricane Sandy on the HRE and observed a dramatic shift in both plankton assemblages and nutrient loads. Cumulatively, these results and the importance of precipitation and river discharge in the RF model suggest *H. akashiwo* bloom events are likely after future major storm events, which is a concerning discovery considering the expected increase in severe storms with climate change (Diffenbaugh et al., 2013; Horton et al., 2015).

Dissolved oxygen appears at several depths and lag times as an important predictor for *H. akashiwo* abundance. Normally, dissolved oxygen is associated with HABs as a result, as

elaborated in the introduction. However the presence as a predictor suggests dissolved oxygen concentrations at certain levels can be a precursor to HABs. Unfortunately, the model fails to elaborate on how dissolved oxygen works as a categorizer (whether high or low concentrations are predictive of high *H. akashiwo* abundance). There is scarce literature on the use of dissolved oxygen levels to predict algal bloom activity as most research focuses on the reverse, using algal bloom activity to predict dissolved oxygen.

While the abiotic factors thus far discussed have reasonably well understood causal relationships to phytoplankton, the symbiotic relationships are less well understood. The first thing to note is that this analysis is correlative, and as such there is potential for both *H. akashiwo* and the identified predictor species to both be responding to some other stimuli with differing lag times. Further, autocorrelation was intentionally unadjusted for in the RF model. Often, models examining correlation between variables in an environmental time series will adjust for autocorrelation as a way of reducing the identification of correlations which may only appear due to trends in seasonality. This increases the likelihood that correlations identified would have potentially meaningful relationships. However, autocorrelation was not adjusted for in the RF model. We reasoned that seasonal correlations between variables had the potential for improving prediction accuracy. This has the potential to add another layer of obscurity to the interpretation of variable importance, however this was a sacrifice made for the sake of model accuracy. Regardless, the biotic variables listed as important predictors are worth investigating.

The two most important biotic variables according to **Figure 5** are *Chlamydomonas spp.* and *Heterocapsa rotundata*. The ecological network analysis publication by Rothenberger et al. (2023) also found a strong, significant, positive correlation between *Chlamydomonas spp.* and *H.*

akashiwo when analyzing the network produced at site 6. Also, *Chlamydomonas spp.* was identified as a predictor for both Lag 1 and Lag 2. *Chlamydomonas spp.* is a genus of green algae which has been present in the HRE throughout the sampling period. Although the exact relationship between *Chlamydomonas spp.* and *H. akashiwo* is unknown due to a lack of literature, both species possess a specialized physiology which enables nitrate reductase to convert nitric oxide into nitrate; *Chlamydomonas spp.* further appears to use nitrate reductase to produce nitric oxide in other conditions (Healey et al., 2023). The former mechanism suggests both species may do well in environments where nitric oxide is more available than other forms of nitrogen while the latter suggests *Chlamydomonas spp.* may play some commensal role by producing nitric acid for *H. akashiwo*. Either of these mechanisms would support a positive relationship between the two taxa, but further experimental research is required to support this.

The relationship between *H. rotundata* and *H. akashiwo* was not reflected in the ecological network analysis, however a relationship between these variables may still exist. Both species are known HAB producers in the HRE so it is possible that they may both respond to similar bloom-initiating stimuli with different lag times, or that there is a competition aspect between them where *H. rotundata* uses the resources before *H. akashiwo*, and they exhibit a negative correlation. Lemley et al. (2018) studied both species and their HAB activity in the Sundays Estuary of South Africa and observed *H. rotundata* was most abundant in winter months but would be suppressed in the presence of *H. akashiwo*, likely due to photosynthetic efficiency and nutrient uptake competitive advantages. Due to the nature of the model, we cannot tell if the nature of the predictive relationship between the variables, however it warrants further investigation in future studies.

4.3. Water Quality of the HRE

Summary of a few key water quality parameters was performed in this study in an attempt to get a general idea of how the HRE has changed over our long-term study. It is clear that the system continues to experience environmental degradation. The decrease in dissolved oxygen is particularly concerning as fish species require certain threshold levels to maintain healthy populations. The standard general for estuarine dissolved oxygen is 4 mg/L as set by the New Jersey Department of Environmental Protection (*Dissolved Oxygen in Coastal Waters*, 2021). Although the mid-summer seasonal lows in dissolved oxygen within our dataset have not yet been observed to violate this standard, it is clear that continuation of current trends will soon lead to regular summer violations.

Two mechanisms are likely candidates for explaining this trend. First, the increasing frequency and intensity of algal bloom activity, as suggested by Anderson et al. (2021), may be transpiring within the HRE, although we lack additional data to support this. Another likely explanation is rising global temperatures which increase water temperature and thereby reduce the solubility of oxygen in water. However, trends in water temperature over the course of the study failed to show significance (p-value > 0.3 for all depths).

Reduction in pH also was observed, consistent with ocean acidification as a result of climate change. Acidification may have different influences between phytoplankton taxa, however *H. akashiwo* has shown greater success in natural waters with higher concentrations of CO_2 (Fu et al., 2008). Microcosm research based on the HRE provides further evidence of *H.*

akashiwo success in acidified waters (Gleich, 2017). Other experiments similarly have found decreasing pH stimulates the growth of *H. akashiwo* and proposes this may be related to nutrient availability, particularly of ferrous iron (Matheson, 2014). In systems where ferric iron may accumulate, ocean acidification will begin to reduce ferric iron to ferrous, becoming readily available for species like *H. akashiwo* which utilize the nutrient. The acidification trends observed resulting from climate change throughout the HRE may be playing a role in the frequent HAB formation of the species. Cumulatively this is a concerning dynamic of climate change that may worsen HAB issues on a global scale.

The significant change in ammonium observed also causes concern for the health of the HRE. Nitrogen is widely recognized as a limiting nutrient for phytoplankton growth, and influx of nutrients like ammonium into aquatic ecosystems are often responsible for eutrophication and destabilization of phytoplankton assemblages, often resulting in algal blooms (Hecky & Kilham, 1988; Livingston, 2001). However the relationship of ammonium with algal blooms can be complicated, as phytoplankton abundances can be stimulated in low-nitrate systems or conversely suppressed in systems where nitrate is high (Glibert et al., 2016). The influx of ammonium into the HRE is likely an aggregate of several sources including mineralization of organic nitrogen by microorganisms, reduction of nitrate through dissimilatory nitrate reduction, fertilizer/manure runoff, and sewage/wastewater (Covatti & Grischek, 2021). It seems unlikely that natural biotic, geologic and hydrologic processes would explain the strong significance and rate of change detected in the ammonium trend analysis, and as such it reasons that anthropogenic drivers may be responsible. Several wastewater treatment facilities in New York and New Jersey have been identified as polluters of the HRE in past studies and could explain

some of the change (M. B. Rothenberger et al., 2014). Regardless of the true explanation for the trend, the influx of nitrogen is likely to continue stimulating algal blooms in the near future.

The change in Secchi depth reflects an increase in water clarity, generally atypical for systems experiencing eutrophication where a general symptom of eutrophication is decreased light availability (Livingston, 2001). According to the Theil-Sen's slope, the magnitude is about 0.12m per year. This is very minimal and generally Secchi depth for the HRE is small compared to other estuaries, a symptom of its eutrophic condition. The depths exhibited by the HRE are notably consistent with those of the Chesapeake Bay which is generally considered a eutrophic system in poor health. Chesapeake Bay Secchi depths are regularly less than 2 meters according to an analysis of a dataset with observations from 1960-2015 (Harding et al., 2019). This compared with more pristine estuaries like Kachimak Bay in Alaska where the lowest Secchi depth values were between 2-4 meters and were as great as 11 meters in some areas (Hartwell et al., 2009).

Additionally, the increasing trend of water clarity in the HRE determined by a linear model may fail to represent the whole story. Clarity seems to be consistent between 2010-2014, but then experiences a major increase from 2014-2016, followed by a gradual return to normal levels. The reason for this fluctuation is unknown at this time.

4.4. Broader Applications

As Hallegraff et al. (2021) point out, the diverse array of aquatic ecosystems make application of knowledge in one difficult to apply directly to another, and therefore site-by-site analysis is necessary for each ecosystem. Our long-term study on the HRE is the only known

long-term monitoring performed on this ecosystem in the past several decades. Application of the RF model developed in this paper has potential for management as mentioned earlier in the discussion, but only for *H. akashiwo* in the HRE.

However, the RF methodology used in this paper appears to be effective based on built-in evaluation methods, even with monthly sampling. This method can be applied to many other systems which are more closely monitored. RF models provide an advantage over more standard correlative techniques in their predictive ability and their capability to impose conditional variables on relationships. For example, while correlation can determine the relationship between X and Y, the random forest can find a more hidden relationship between X and Y when Z possesses a certain condition. The consequence of this is the nature of the relationships are highly obscured and not easily interpreted from a biological perspective. Further, the predictive style model allows users to forecast future values, opposed to inference style models which are more concerned with unearthing the relationships between variables which would be highly challenging with our complex observational dataset. However the evaluation of important variables in the predictive model serves as a hypothesis generation tool for future experimental research.

Finally, we especially recommend the use of weighted observations as a method of improving accuracy and precision while decreasing error, especially for datasets with lower "high" to "low" abundance observation ratios. This method has not been observed in other RF HAB models. Additionally, evaluation of different parameters by permutation of all combinations is highly recommended since the typical values for RF parameters are not always optimal for the dataset, as was the case for our study. Utilization of these methods for datasets in

other systems and evaluating blooms of other species have potential to yield even more information on the complex nature of algal blooms and phytoplankton communities.

Appendix I: Figures

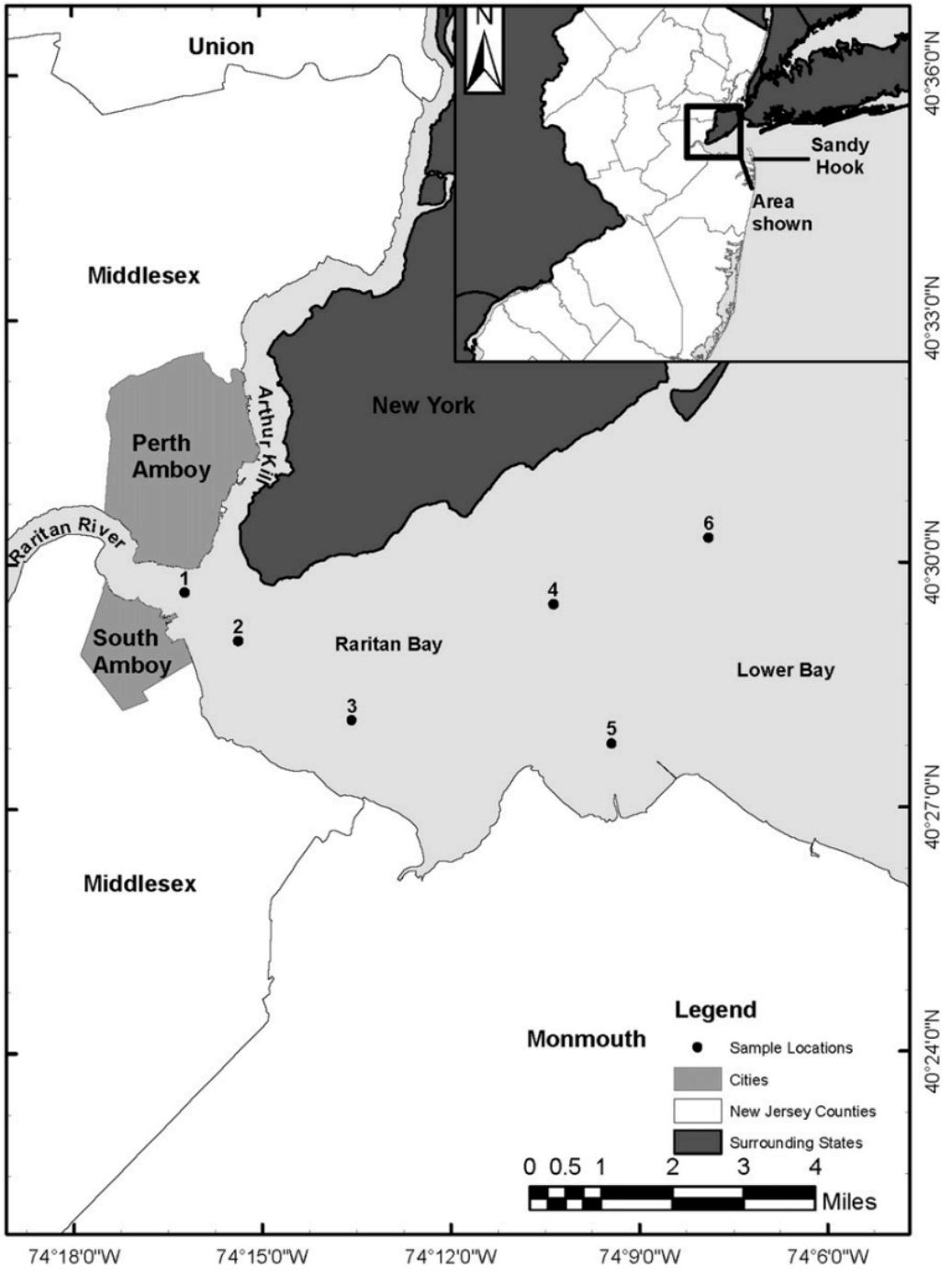

Figure 1. Map of the Hudson-Raritan Estuary with sampling locations. Figure borrowed from (M. B. Rothenberger & Calomeni, 2016).

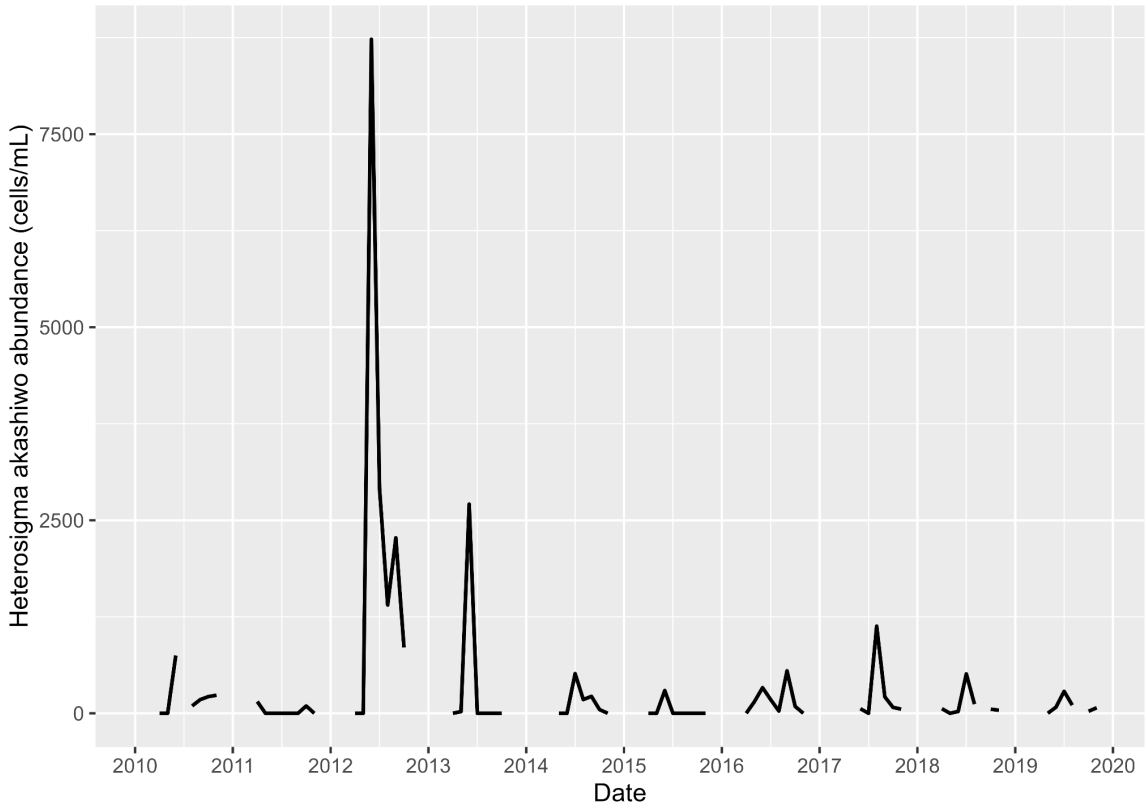

Figure 2. *Heterosigma akashiwo* abundance 2010-2020. For consecutive monthly samples, data points are connected, gaps indicate absent data for at least one month of sample.

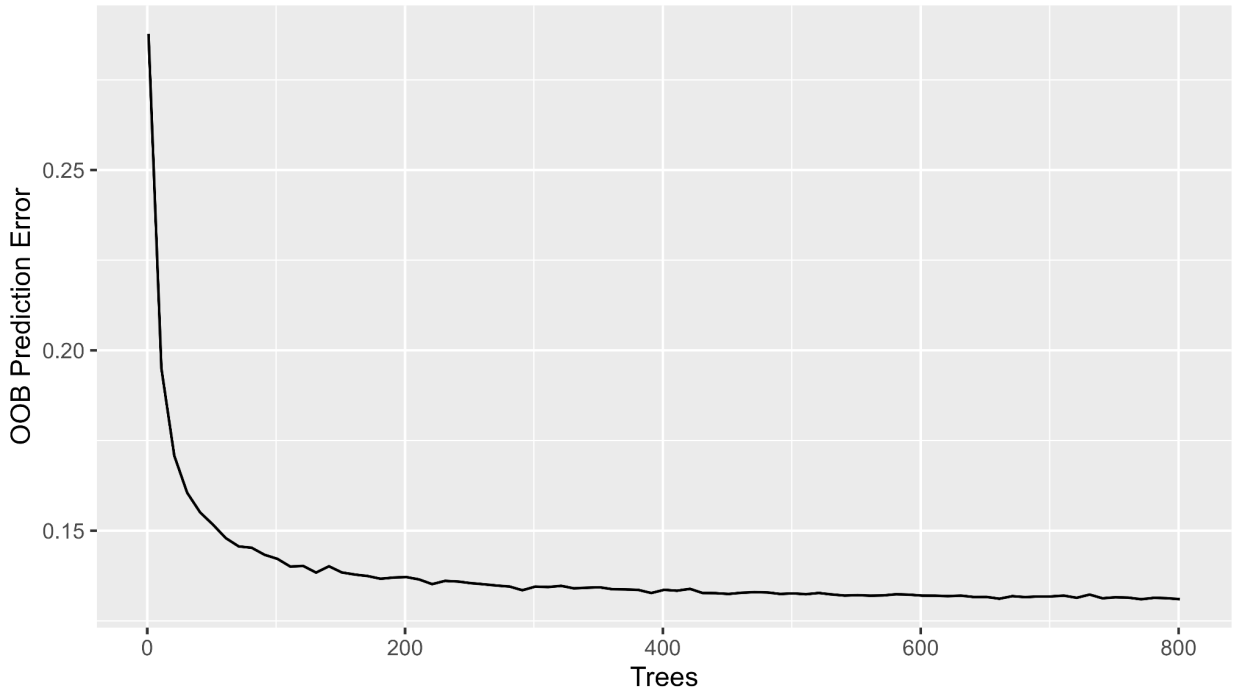

Figure 3. Average OOB prediction error for all variable combinations vs number of trees used in random forest construction.

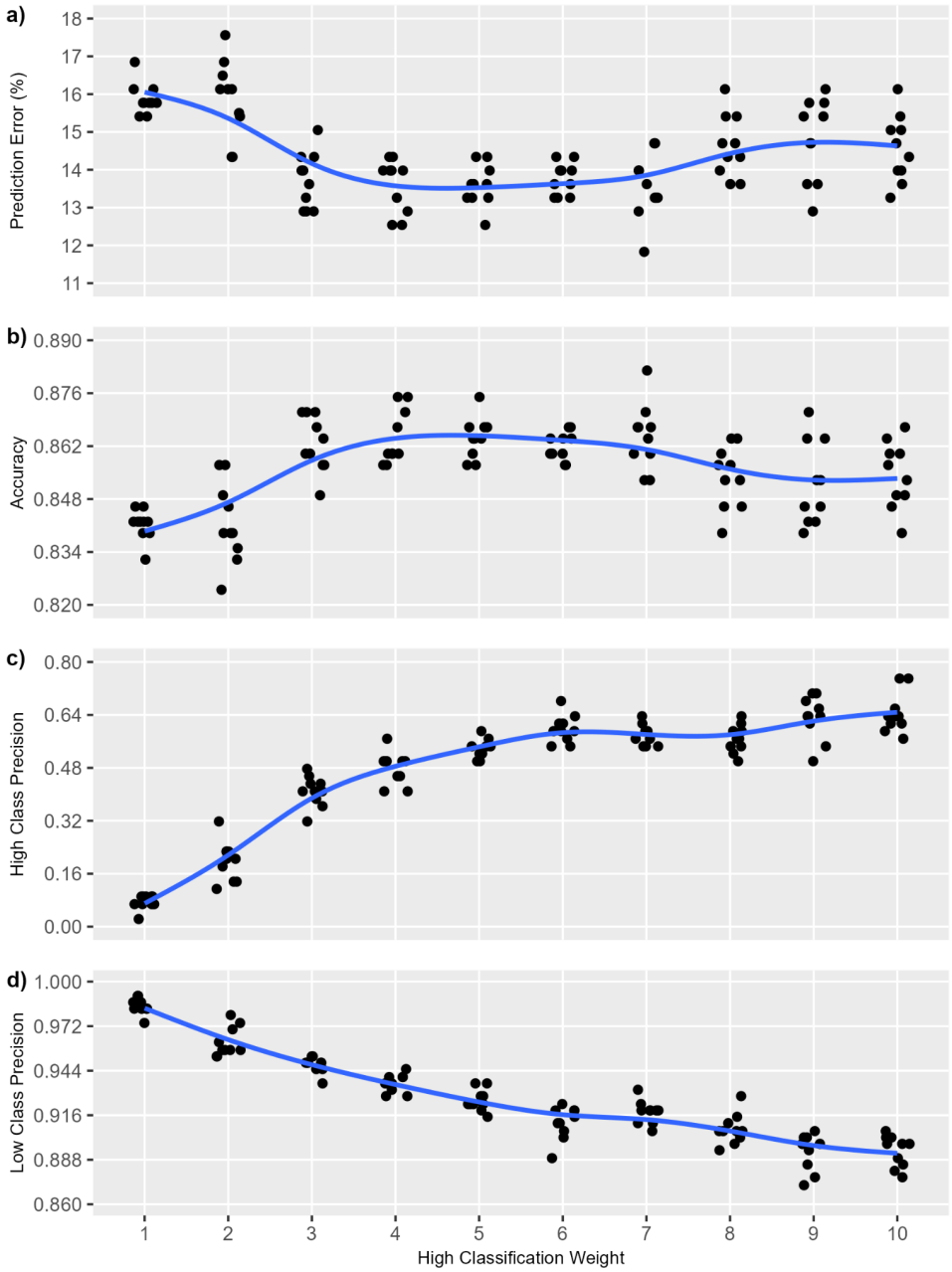

Figure 4. Plots showing the costs and benefits of a weighting parameter evaluated my a) OOB error, b) accuracy, c) precision for "high" classification, and d) precision for "low" classifications. Ten RF models were generated and evaluated for optimal parameters (**Table 1**) for each weighting scheme, prediction error was available in the RF model summary and values for b), c), and d) were calculated from the associated confusion matrices.

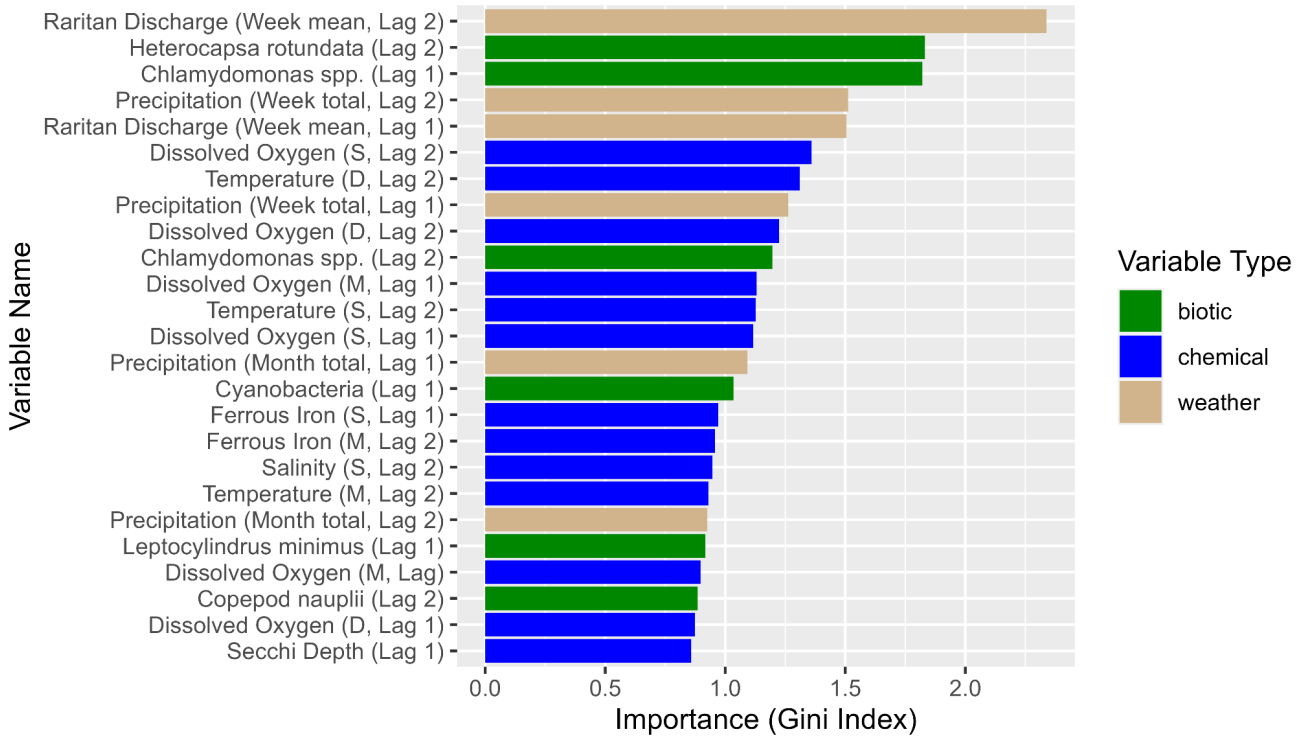

Figure 5. Variable importance plot for the optimal RF model colored by variable category. Biotic variables refer to phytoplankton- or zooplankton-related variables, weather refers to precipitation and discharge variables, and chemical refers to any geochemical water variables.

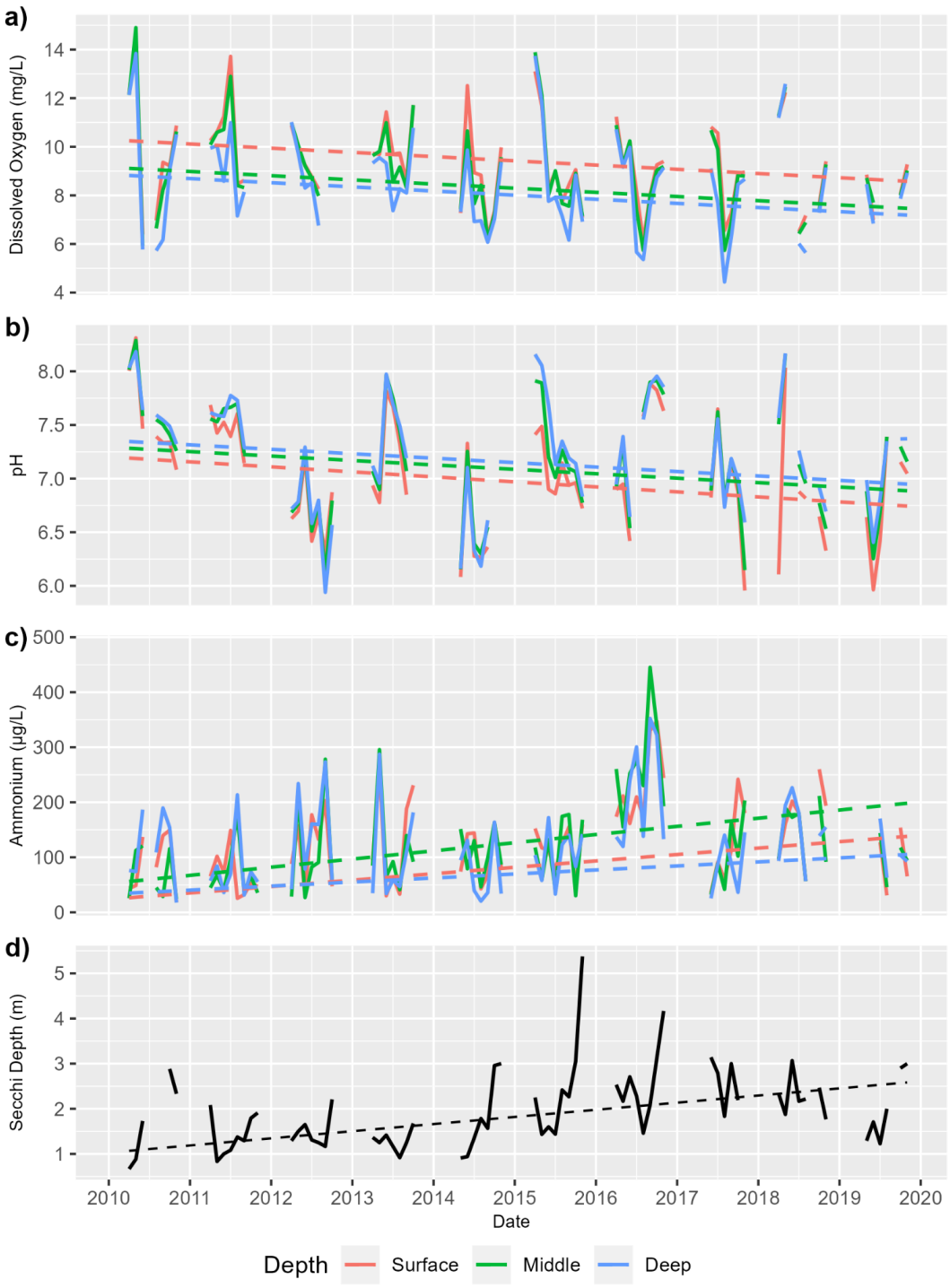

Figure 6. Time series plots for a) dissolved oxygen, b) pH, c) ammonium, and d) Secchi depth. Trends are displayed for variables which showed significant change upon evaluation using Seasonal Mann-Kendall test (data in **Table 4.**).

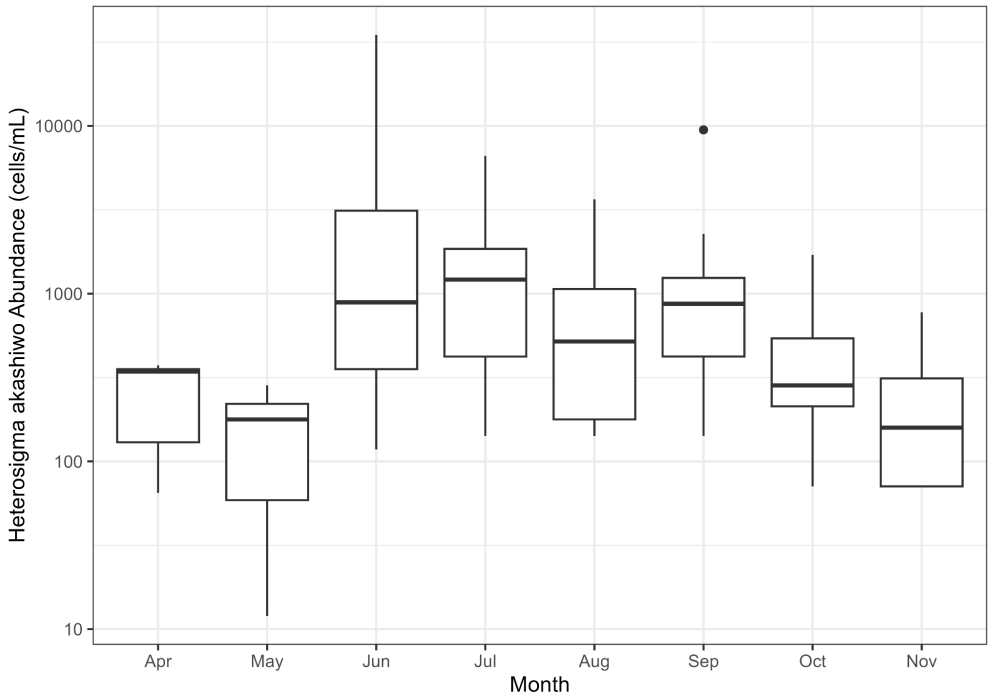

Figure 7. Seasonal variation in *H. akashiwo* illustrated by box plots. Samples were collected from surface water (depth = 0.5 m) for the months of April-November over a 12 year sampling period.

Appendix II: Tables

Table 1. Parameter values tested in development of RF model. All combinations of values were evaluated using OOB error, bold values are those closest to suggested values p□ and p□ for typical RF construction.

Variables Tested Per Node (p_t)	Minimum Node Size (p_n)	Percent of Observations in Bootstrap (p_s)
15	2	0.45
17	3	0.50
19	4	0.55
21	5	0.60
23	6	**0.65**
25	7	0.70
27	8	0.75
29	9	0.80

Table 2. Evaluation of ideal weighting for "high" algal bloom density in RF model development. All values were calculated as the mean of 10 trials with variance in parentheses.

		Confusion Matrix Statistics		
Weight	OOB Prediction Error (%)	Accuracy	Precision ("High" Classification)	Precision ("Low" Classification)
1 (no weighting)	15.88	0.841	0.075	0.985
2	15.89	0.842	0.198	0.962
3	13.73	0.863	0.409	0.948
4	13.58	0.864	0.480	0.936
5	13.58	0.864	0.536	0.926
6	13.76	0.862	0.598	0.911
7	13.55	0.865	0.580	0.918
8	13.59	0.864	0.580	0.917
9	13.62	0.864	0.582	0.917
10	13.80	0.862	0.580	0.915

Table 3. Confusion matrix for the optimal RF model. Observations were classified according to actual classification and classification predicted by the independently constructed trees in the RF model.

		Predicted	
		High	Low
Observed	High	28	16
	Low	17	218

Table 4. Seasonal Mann-Kendall statistics, p-values, and Theil-Sen's slope calculated for surface (0.5 m below surface), middle (0.15 m below surface), and deep (3.0 m below surface). Bold values with stars indicate significance at $\alpha = 0.05$.

Variable	Sample Depth	Tau Statistic	p-value	Theil-Sen Slope	Units
Dissolved Oxygen	Surface	-0.277	**0.0058***	-0.19	mg/L
	Middle	-0.285	**0.0045***	-0.19	mg/L
	Deep	-0.261	**0.0095***	-0.18	mg/L
pH	Surface	-0.216	**0.0325***	-0.07	---------
	Middle	-0.216	**0.0325***	-0.06	---------
	Deep	-0.200	**0.0481***	-0.05	---------
Nitrate	Surface	0.110	0.2589	22.50	µg/L
	Middle	0.014	0.9126	0.00	µg/L
	Deep	0.007	0.9709	0.00	µg/L
Ammonium	Surface	0.255	**0.0083***	8.13	µg/L
	Middle	0.291	**0.0026***	9.31	µg/L
	Deep	0.113	0.2486	4.32	µg/L
SRP	Surface	-0.065	0.5124	-1.08	µg/L
	Middle	-0.076	0.4444	-0.78	µg/L
	Deep	-0.093	0.3440	-0.69	µg/L
N:P	Surface	0.163	0.0940	1.43	---------
	Middle	0.135	0.1685	0.83	---------
	Deep	0.106	0.2805	0.77	---------
Si:N	Surface	-0.078	0.4345	-0.20	---------
	Middle	-0.070	0.4929	-0.05	---------
	Deep	0.021	0.8524	0.03	---------
Salinity	Surface	-0.099	0.3212	-0.12	ppt
	Middle	-0.040	0.7032	-0.04	ppt
	Deep	-0.040	0.7032	-0.05	ppt
Secchi Depth	N/A	0.415	**< 0.0001***	0.12	m

Acknowledgements

I would like to thank Dr Megan Rothenberger, Dr Trent Gaugler, and Dr Allison Lewis for advising my work and providing guidance through this project. Also thank you to Olivia Sterentino, Declan Winters, Charles Kelshaw, Madison Lebish, and Grace Harvey for their assistance as well as all current and past members of Dr Rothenberger's lab for collecting data over the past 12 years which was used in this project. Thanks to Lafayette College for providing funding for my project.

Citations

Anderson, D. M., Fensin, E., Gobler, C. J., Hoeglund, A. E., Hubbard, K. A., Kulis, D. M., Landsberg, J. H., Lefebvre, K. A., Provoost, P., Richlen, M. L., Smith, J. L., Solow, A. R., & Trainer, V. L. (2021). Marine harmful algal blooms (HABs) in the United States: History, current status and future trends. *Harmful Algae*, *102*, 101975. https://doi.org/10.1016/j.hal.2021.101975

Behrenfeld, M. J., Bale, A. J., Kolber, Z. S., Aiken, J., & Falkowski, P. G. (1996). Confirmation of iron limitation of phytoplankton photosynthesis in the equatorial Pacific Ocean. *Nature*, *383*(6600), 508–511.

Berg, D. L., & Levinton, J. S. (1985). *The biology of the Hidson-Raritan Estuary, with emphasis on fishes*. NOAA.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breitburg, D. L., Pihl, L., & Kolesar, S. E. (2001). Effects of low dissolved oxygen on the behavior, ecology and harvest of fishes: A comparison of the Chesapeake Bay and Baltic-Kattegat Systems. In N. N. Rabalais & R. E. Turner (Eds.), *Coastal and Estuarine Studies* (Vol. 58, pp. 241–267). American Geophysical Union. https://doi.org/10.1029/CE058p0241

Breteler, R. J. (1984). *Chemical pollution of the Hudson-Raritan estuary*.

Burkholder, J. M., Dickey, D. A., Kinder, C. A., Reed, R. E., Mallin, M. A., McIver, M. R., Cahoon, L. B., Melia, G., Brownie, C., Smith, J., Deamer, N., Springer, J., Glasgow, H.

B., & Toms, D. (2006). Comprehensive trend analysis of nutrients and related variables in a large eutrophic estuary: A decadal study of anthropogenic and climatic influences. *Limnology and Oceanography*, *51*(1part2), 463–487. https://doi.org/10.4319/lo.2006.51.1_part_2.0463

Burkholder, J. M., Tomasko, D. A., & Touchette, B. W. (2007). Seagrasses and eutrophication. *Journal of Experimental Marine Biology and Ecology*, *350*(1–2), 46–72. https://doi.org/10.1016/j.jembe.2007.06.024

CDC | Case Definition: Brevetoxin Poisoning. (2019, May 15). https://emergency.cdc.gov/agent/brevetoxin/casedef.asp#print

Cloern, J. E., Foster, S. Q., & Kleckner, A. E. (2014). Phytoplankton primary production in the world's estuarine-coastal ecosystems. *Biogeosciences*, *11*(9), 2477–2501. https://doi.org/10.5194/bg-11-2477-2014

Covatti, G., & Grischek, T. (2021). Sources and behavior of ammonium during riverbank filtration. *Water Research*, *191*, 116788. https://doi.org/10.1016/j.watres.2020.116788

Derot, J., Yajima, H., & 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*, 101906. https://doi.org/10.1016/j.hal.2020.101906

Diffenbaugh, N. S., Scherer, M., & Trapp, R. J. (2013). Robust increases in severe thunderstorm environments in response to greenhouse forcing. *Proceedings of the National Academy of Sciences*, *110*(41), 16361–16366. https://doi.org/10.1073/pnas.1307758110

Dissolved Oxygen in Coastal Waters. (2021). [Environmental Trends Report]. NJDEP.

Driscoll, W. W., Hackett, J. D., & Ferrière, R. (2016). Eco‐evolutionary feedbacks between

private and public goods: Evidence from toxic algal blooms. *Ecology Letters*, *19*(1), 81–97. https://doi.org/10.1111/ele.12533

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, *281*(5374), 237–240. https://doi.org/10.1126/science.281.5374.237

Fu, F.-X., Zhang, Y., Warner, M. E., Feng, Y., Sun, J., & Hutchins, D. A. (2008). A comparison of future increased CO2 and temperature effects on sympatric Heterosigma akashiwo and Prorocentrum minimum. *Harmful Algae*, *7*(1), 76–90. https://doi.org/10.1016/j.hal.2007.05.006

Gleich, S. J. (2017). *Interactive effects of iron enrichment, zooplankton grazing and acidification on coastal phytoplankton assemblages*. Lafayette College.

Glibert, P. M., Wilkerson, F. P., Dugdale, R. C., Raven, J. A., Dupont, C. L., Leavitt, P. R., Parker, A. E., Burkholder, J. M., & Kana, T. M. (2016). Pluses and minuses of ammonium and nitrate uptake and assimilation by phytoplankton and implications for productivity and community composition, with emphasis on nitrogen-enriched conditions: Pluses and minuses of NH4+ *and* NO3−. *Limnology and Oceanography*, *61*(1), 165–197. https://doi.org/10.1002/lno.10203

Graham, L. E., & Wilcox, L. W. (2000). *Algae*. Prentice Hall.

Griffin, S. L., Herzfeld, M., & Hamilton, D. P. (2001). Modeling the impact of zooplankton grazing on phytoplankton biomass during a dinoflagellate bloom in the Swan River Estuary, western Australia. *Ecological Engineering*, *16*(3), 373–394. https://doi.org/10.1016/S0925-8574(00)00122-1

Hallegraeff, G. M., Anderson, D. M., Belin, C., Bottein, M.-Y. D., Bresnan, E., Chinain, M., Enevoldsen, H., Iwataki, M., Karlson, B., McKenzie, C. H., Sunesen, I., Pitcher, G. C., Provoost, P., Richardson, A., Schweibold, L., Tester, P. A., Trainer, V. L., Yñiguez, A. T., & Zingone, A. (2021). Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. *Communications Earth & Environment*, *2*(1), 117. https://doi.org/10.1038/s43247-021-00178-8

Harding, L. W., Mallonee, M. E., Perry, E. S., Miller, W. D., Adolf, J. E., Gallegos, C. L., & Paerl, H. W. (2019). Long-term trends, current status, and transitions of water quality in Chesapeake Bay. *Scientific Reports*, *9*(1), 6709. https://doi.org/10.1038/s41598-019-43036-6

Harley, J. R., Lanphier, K., Kennedy, E., Whitehead, C., & Bidlack, A. (2020). Random forest classification to determine environmental drivers and forecast paralytic shellfish toxins in southeast Alaska with high temporal resolution. *Harmful Algae*, *99*, 101918. https://doi.org/10.1016/j.hal.2020.101918

Hartwell, S. I., Apeti, A. D., Claflin, W. L., Johnson, W. E., & Kimbrough, L. K. (2009). *Sediment quality triad assessment in Kachemak Bay: Characterization of soft bottom benthic habitats and contaminant bioeffects assessment*.

Healey, E. M., Flood, S., Bock, P. K., Fulweiler, R. W., York, J. K., & Coyne, K. J. (2023). Effects of nitrate and ammonium on assimilation of nitric oxide by Heterosigma akashiwo. *Scientific Reports*, *13*(1), 621. https://doi.org/10.1038/s41598-023-27692-3

Hecky, R. E., & Kilham, P. (1988). Nutrient limitation of phytoplankton in freshwater and marine environments: A review of recent evidence on the effects of enrichment: nutrient

enrichment. *Limnology and Oceanography*, *33*(4part2), 796–822. https://doi.org/10.4319/lo.1988.33.4part2.0796

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, *18*(1), 107–121. https://doi.org/10.1029/WR018i001p00107

Horton, R., Bader, D., Kushnir, Y., Little, C., Blake, R., & Rosenzweig, C. (2015). New York City Panel on Climate Change 2015 ReportChapter 1: Climate Observations and Projections: NPCC 2015 Report Chapter 1. *Annals of the New York Academy of Sciences*, *1336*(1), 18–35. https://doi.org/10.1111/nyas.12586

Hudson River Foundation. (2022). https://www.hudsonriver.org/estuary-program

Hudson-Raritan Estuary Ecosystem Restoration Feasibility Study. (2020). [Final Integrated Feasibility Report and Environmental Assessment]. U.S. Army Corps of Engineers.

Hutchins, D. A., DiTullio, G. R., Zhang, Y., & Bruland, K. W. (1998). An iron limitation mosaic in the California upwelling regime. *Limnology and Oceanography*, *43*(6), 1037–1054. https://doi.org/10.4319/lo.1998.43.6.1037

Hutchinson, G. (1961). The paradox of the plankton. *The American Naturalist*, *95*(883), 137–145.

Jeffries, H. P. (1962). Environmental characteristics of Raritan Bay, a polluted estuary. *Limnology and Oceanography*, *7*(1), 21–31. JSTOR.

Jin, D., & Hoagland, P. (2008). The value of harmful algal bloom predictions to the nearshore commercial shellfish Fishery in the Gulf of Maine. *Harmful Algae*, *7*(6), 772–781. https://doi.org/10.1016/j.hal.2008.03.002

Kane, R., & Kerlinger, P. (1994). Raritan Bay wildlife habitat report with recommendations for conservation. *Bernardsville: New Jersey Audubon Society*.

Keppler, C. J., Hoguet, J., Smith, K., Ringwood, A. H., & Lewitus, A. J. (2005). Sublethal effects of the toxic alga Heterosigma akashiwo on the southeastern oyster (Crassostrea virginica). *Harmful Algae*, *4*(2), 275–285. https://doi.org/10.1016/j.hal.2004.05.002

Khan, S., Arakawa, O., & Onoue, Y. (1997). Neurotoxins in a toxic red tide of Heterosigma akashiwo (Raphidophyceae) in Kagoshima Bay, Japan. *Aquaculture Research*, *28*(1), 9–14. https://doi.org/10.1046/j.1365-2109.1997.t01-1-00823.x

Legrand, C., Rengefors, K., Fistarol, G. O., & Granéli, E. (2003). Allelopathy in phytoplankton—Biochemical, ecological and evolutionary aspects. *Phycologia*, *42*(4), 406–419. https://doi.org/10.2216/i0031-8884-42-4-406.1

Lemley, D. A., Adams, J. B., & Rishworth, G. M. (2018). Unwinding a tangled web: A fine-scale approach towards understanding the drivers of harmful algal bloom species in a eutrophic South African estuary. *Estuaries and Coasts*, *41*, 1356–1369.

Livingston, R. J. (2001). *Eutrophication processes in coastal systems: Origin and succession of plankton blooms and effects on secondary production in Gulf Coast estuaries*. CRC Press.

Matheson, J. R. (2014). *The effects of ocean acidification and eutrophication on the growth, lipid composition and toxicity of the marine raphidophyte Heterosigma akashiwo.* (2701132764) [The University of Western Ontario]. ProQuest Dissertations & Theses A&I. https://ezproxy.lafayette.edu/login?url=https://www.proquest.com/dissertations-theses/eff

ects-ocean-acidification-eutrophication-on/docview/2701132764/se-2

Meals, D. W., Spooner, J., Dressing, S. A., & Harcum, J. B. (2011). *Statistical analysis for monotonic trends* (No. 6; Tech Notes). Environmental Protectino Agency by Tetra Tech, Inc. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoringtechnical-notes

Paulsson, O., & Widerlund, A. (2021). Algal nutrient limitation and metal uptake experiment in the Åkerberg pit lake, northern Sweden. *Applied Geochemistry*, *125*, 104829. https://doi.org/10.1016/j.apgeochem.2020.104829

Pettersson, L. H., & Pozdni͡akov, D. V. (2013). *Monitoring of harmful algal blooms*. Springer, published in association with Praxis Publishing.

Rantajärvi, E., Olsonen, R., Hällfors, S., Leppänen, J.-M., & Raateoja, M. (1998). Effect of sampling frequency on detection of natural variability in phytoplankton: Unattended high-frequency measurements on board ferries in the Baltic Sea. *ICES Journal of Marine Science*, *55*(4), 697–704. https://doi.org/10.1006/jmsc.1998.0384

Reynolds, C. S. (2006). *Ecology of phytoplankton*. Cambridge University Press.

Rice, E. W., Bridgewater, L., American Public Health Association, American Water Works Association, & Water Environment Federation. (2012). *Standard Methods for the Examination of Water and Wastewater*. American Public Health Association. https://books.google.com/books?id=dd2juAAACAAJ

Rothenberger, M., Armstrong, A., & Spitz, M. (2018). Social–ecological system responses to Hurricane Sandy in the Hudson-Raritan Estuary. *Ambio*, *47*(3), 284–297.

https://doi.org/10.1007/s13280-017-0949-z

Rothenberger, M. B., & Calomeni, A. J. (2016). Complex interactions between nutrient enrichment and zooplankton in regulating estuarine phytoplankton assemblages: Microcosm experiments informed by an environmental dataset. *Journal of Experimental Marine Biology and Ecology*, *480*, 62–73. https://doi.org/10.1016/j.jembe.2016.03.015

Rothenberger, M. B., Swaffield, T., Calomeni, A. J., & Cabrey, C. D. (2014). Multivariate analysis of water quality and plankton assemblages in an urban estuary. *Estuaries and Coasts*, *37*(3), 695–711. https://doi.org/10.1007/s12237-013-9714-0

Rothenberger, M., Gleich, S. J., & Flint, E. (2023). The underappreciated role of biotic factors in controlling the bloom ecology of potentially harmful microalgae in the Hudson-Raritan Bay. *Harmful Algae*, 102411. https://doi.org/10.1016/j.hal.2023.102411

Sarkar, S. K. (2018). *Marine algal bloom: Characteristics, causes and climate change impacts*. Springer Berlin Heidelberg.

Schoenberg, S. A., & Carlson, R. E. (1984). Direct and indirect effects of zooplankton grazing on phytoplankton in a hypereutrophic lake. *Oikos*, *42*(3), 291. https://doi.org/10.2307/3544397

Seymour, J. R., Amin, S. A., Raina, J.-B., & Stocker, R. (2017). Zooming in on the phycosphere: The ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology*, *2*(7), 17065. https://doi.org/10.1038/nmicrobiol.2017.65

Shaked, Y., & Lis, H. (2012). Disassembling Iron Availability to Phytoplankton. *Frontiers in Microbiology*, *3*. https://doi.org/10.3389/fmicb.2012.00123

Sheng, J., Malkiel, E., Katz, J., Adolf, J. E., & Place, A. R. (2010). A dinoflagellate exploits

toxins to immobilize prey prior to ingestion. *Proceedings of the National Academy of Sciences*, *107*(5), 2082–2087. https://doi.org/10.1073/pnas.0912254107

Suthers, I. M., & Rissik, D. (Eds.). (2009). *Plankton: A guide to their ecology and monitoring for water quality*. CSIRO Pub.

Yajima, H., & Derot, J. (2018). Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics*, *20*(1), 206–220. https://doi.org/10.2166/hydro.2017.010

Yamochi, S. (1983). Mechanisms for outbreak ofHeterosigma akashiwo red tide in Osaka Bay, Japan: Part 1. Nutrient factors involved in controlling the growth ofHeterosigma akashiwo Hada. *Journal of the Oceanographical Society of Japan*, *39*(6), 310–316. https://doi.org/10.1007/BF02071827

Yang, Y., Gao, Y., Huang, X., Ni, P., Wu, Y., Deng, Y., & Zhan, A. (2019). Adaptive shifts of bacterioplankton communities in response to nitrogen enrichment in a highly polluted river. *Environmental Pollution*, *245*, 290–299. https://doi.org/10.1016/j.envpol.2018.11.002

Zhu, J., Hong, Y., Zada, S., Hu, Z., & Wang, H. (2018). Spatial variability and co-acclimation of phytoplankton and bacterioplankton communities in the Pearl River Estuary, China. *Frontiers in Microbiology*, *9*, 2503. https://doi.org/10.3389/fmicb.2018.02503