Predictive Modeling of Harmful Algal Blooms in the Western Basin of Lake Erie

Introduction:

The Great Lakes comprise 21% of Earth's surface freshwater by volume.[1] They are a vital resource that provides water for consumption, power, recreation, and other uses. The Great Lakes as a whole provide drinking water for 10% of the U.S. population and 30% of the Canadian population.[2] Lake Erie alone provides drinking water to 11 million people.[3] Unfortunately, pollution from human activity and climate change pose threats to the ecological stability of the Great Lakes as a whole, but Lake Erie in particular. Lake Erie is ecologically vulnerable for three key reasons:

1. It is the shallowest of the Great Lakes, which makes it the most susceptible to warming.[4]
2. It receives a large influx of industrial and agricultural runoff that introduces bioavailable nutrients and impacts its water quality.
3. It has been exposed to invasive species, notably zebra mussels and Asian carp.[5]

Lake Erie's water quality has been a concern since the 1960s, and has again become an issue in the 2000s as the lake has begun to once again experience Harmful Algal Blooms (HABs) in the summer months. HAB refers to the rapid growth of algae that produce harmful toxins, the most common of which is microcystin. Microcystins can cause liver damage, and are harmful if ingested or, at high enough concentrations, through skin contact.[6] In 2014 the Western Basin of Lake Erie experienced an HAB that caused microcystin concentrations to exceed allowable limits for water purification, and 400,000 area residents were without drinking water for several days.[7] As this incident demonstrates, HABs are incredibly disruptive events economically and ecologically.

The goal of this project is to investigate the relationships between water quality and environmental conditions and the presence of algal blooms using water quality data from the Great Lakes Environmental Research Laboratory. The guiding questions of this project are:

1. What is the typical spatial and temporal profile of nutrients and conditions (phosphorus, nitrogen, oxygen, temperature, etc) in Lake Erie in the absence of an HAB event?
2. Which water quality features are most predictive of an HAB event?
3. What are the capabilities and limitations of a predictive model for microcystin concentration in Lake Erie?

---

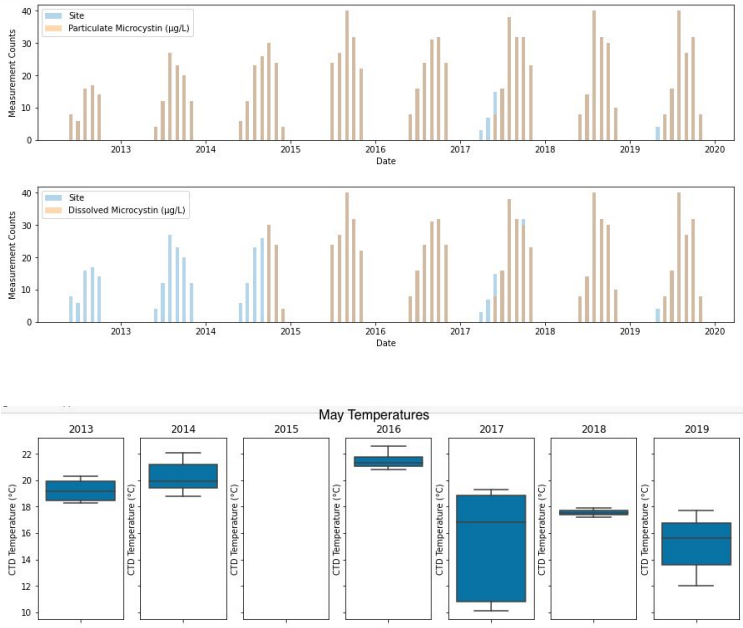
[1] https://www.epa.gov/greatlakes
[2] epa.gov/greatlakes/facts-and-figures
[3] https://phys.org/news/2019-11-lake-erie-people-algae-blooms.html
[4] https://www.epa.gov/greatlakes/lake-erie
[5] https://lakeeriefoundation.org/issues/invasive-species/
[6] https://oehha.ca.gov/media/downloads/ecotoxicology/document/microcystin031209.pdf
[7] https://earthobservatory.nasa.gov/images/84125/algae-bloom-on-lake-erie

Data Wrangling:

The Great Lakes Environmental Research Laboratory (GLERL) and the Cooperative Institute for Great Lakes Research collect water quality data in western Lake Erie in order to monitor the potential for Harmful Algal Blooms (HAB). Beginning in 2012 GLERL began weekly sampling trips to collect water quality data, including microcystin levels, before, during, and after HAB events (from spring to late summer / fall). Some of these collections were performed by boat or automated, movable buoys. Moored buoy station locations were eventually established starting in 2014. The original data can be accessed through NOAA's National Centers for Environmental Information Database[8]. The raw data was downloaded from two repositories - one for archived data and one for current data. Basic cleaning was performed to ensure features were the correct data type and the datasets were merged.[9]

The main challenges to wrangling this data were the spatiotemporal variability in the data and missing values. Spatially, the sample site locations were recorded as a categorical variable. However, based on latitude and longitude analysis the site labels did not reliably correspond with proximate locations from year to year due to ongoing changes in GLERL's sampling methodology. Additionally some samples were recorded as being collected on land. Mapping and filtering based on coordinate boundaries was used to relabel sites consistently so that samples with shared site labels were spatially proximate.

The sampling counts per month were visualized in order to help with feature selection and the determination of null values. Blue bars in the figure represent missing values. Based on this visualization particulate microcystin was identified as the target variable given the several years of missing data for dissolved microcystin. Boxplot visualizations of temperatures for each month demonstrated that there is variability in monthly temperature from year to year, so where possible missing values were imputed using the monthly average for each specific year. Nutrient measurements counts were found to be fairly consistent across time. Further imputation was set aside for feature engineering so that feature distributions could first be explored and understood.



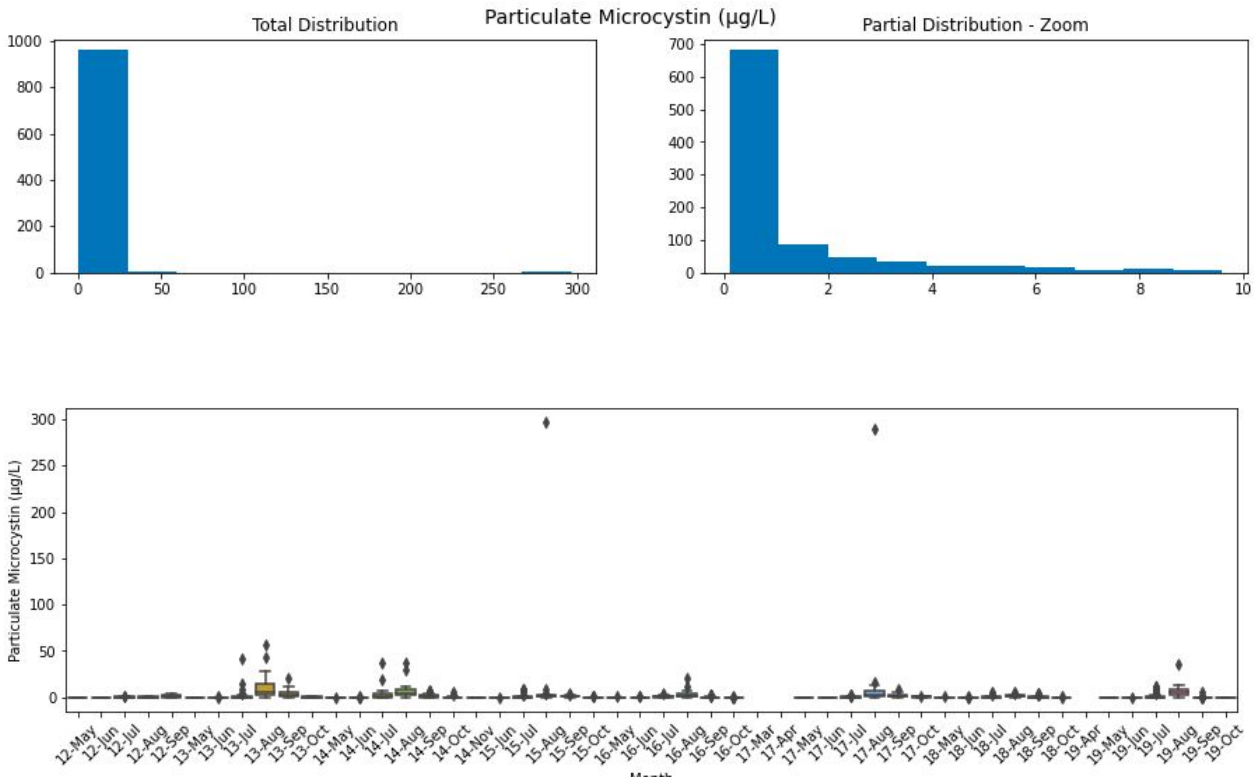
[8]https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:GLERL-CIGLR-HAB-LakeErie-water-qual
[9]https://github.com/melanierbutler/HAB_Capstone/blob/master/notebooks/HAB_Capstone_Data_Wrangling.ipynb

Exploratory Data Analysis: Individual Feature Distributions:

Guiding Question #1: What is the typical spatial and temporal profile of nutrients and conditions (phosphorus, nitrogen, oxygen, temperature, etc) in Lake Erie?

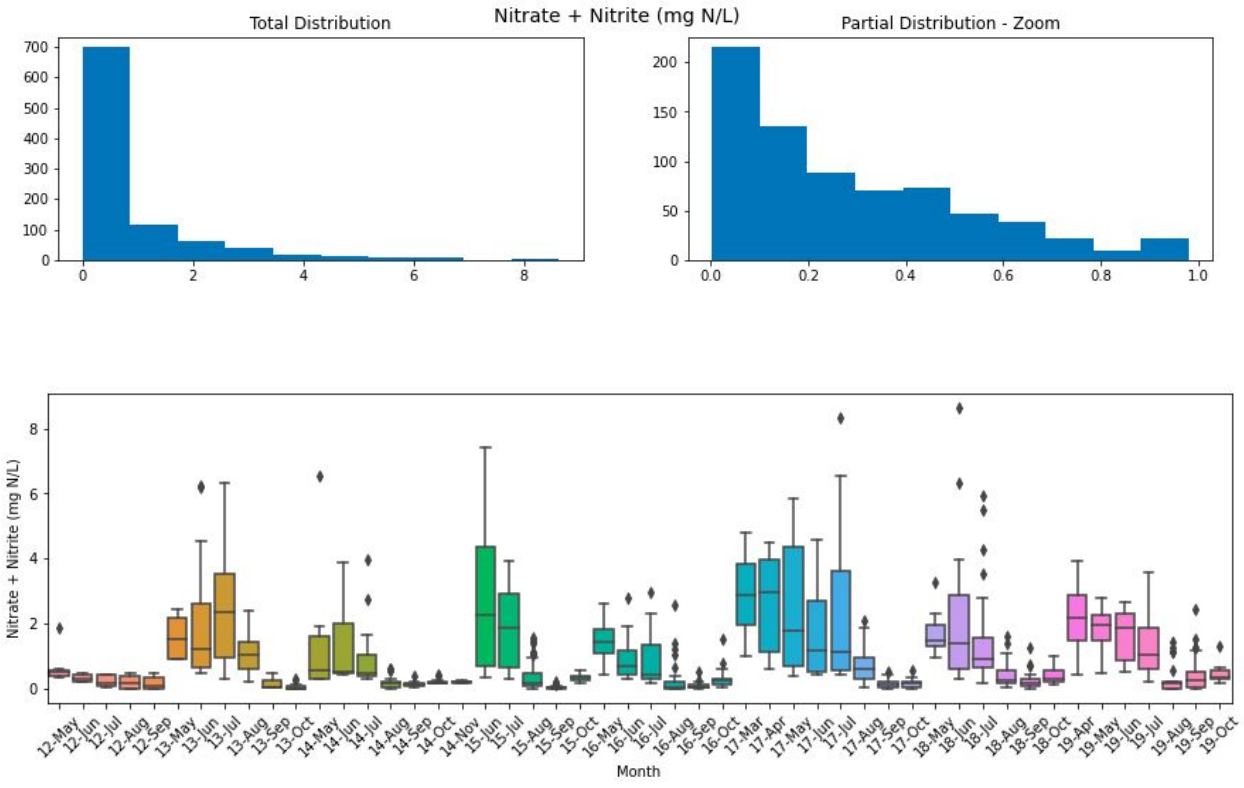
Particulate Microcystin:

Particulate microcystin was identified as the target variable for the predictive model, since this chemical is the primary toxin produced by HABs. While there are many other indicators of algal bloom presence, such as chlorophyll a and phycocyanin, those other indicators can be produced by other organisms. Toxins such as microcystin are one of the main reasons HABs are harmful, so this is a more useful target metric than the other biological features in the data set. Particulate microcystin was found to most often fall between 0 and 1.6 µg/L, which is the Ohio EPA limit for drinking water. There were a fair number of outliers, however, with concentrations up to 300 µg/L. Higher particulate microcystin concentrations reliably occurred in late summer months, mostly August, as visualized by the boxplots below. This is as expected since algae thrive in warmer water with rich nutrient availability.



Nitrogen, Nitrites and Nitrates (NOx):

All forms of nitrogen have a long right tail. In the Great Lakes algal growth is commonly thought to be phosphorus-limited, that is there is usually enough nitrogen to support the start of a bloom, but once the bloom begins and starts consuming the available nitrogen its growth may become nitrogen-limited. There is some debate as to whether reducing nitrogen levels in Lake Erie would be beneficial. UMass Amherst[10] suggests that if inorganic forms of nitrogen, including nitrate and ammonium (NH4+) exceed 0.3 mg/l in the spring then there is sufficient nitrogen to support a bloom come summer. We can clearly see there is a long tail of nitrate and nitrite above 0.3 mg/L.

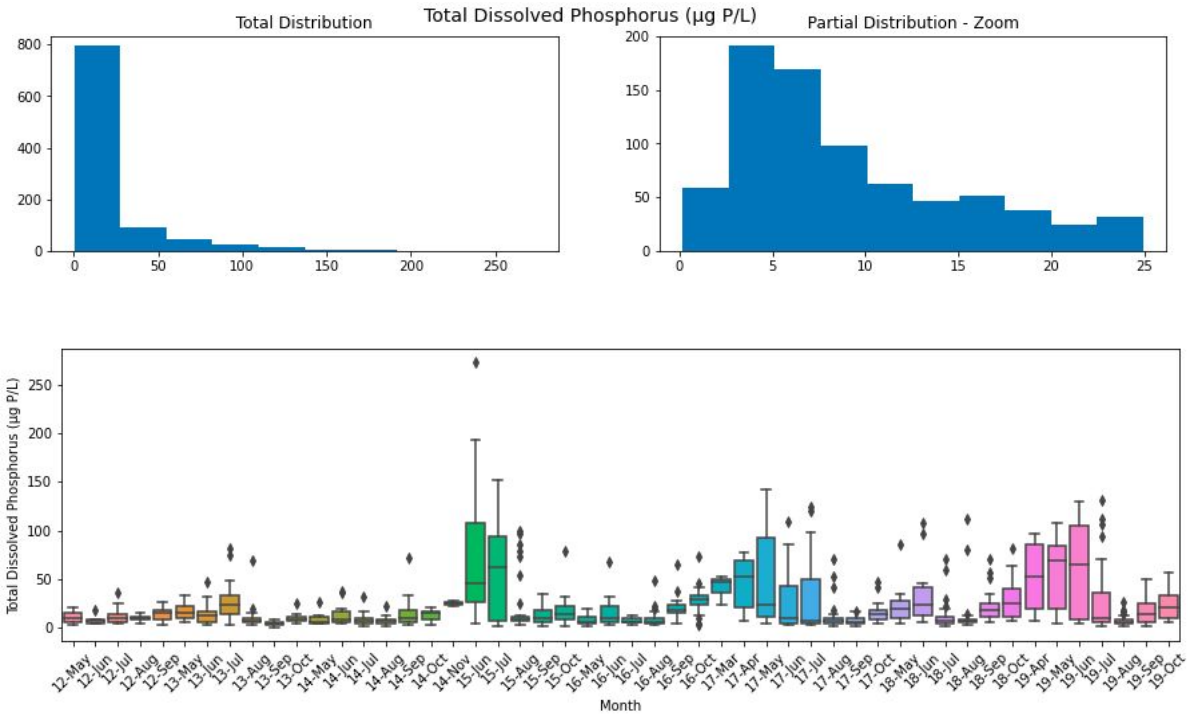


Nitrogen compounds interestingly tend to have higher concentrations and greater variability in the spring months. This may be partly due to greater rainfall and runoff of agricultural pollutants, or increased river loading. This could also be consistent with the notion that as algae and other organisms grow in the warmer months they consume more and more nitrogenous compounds causing the concentrations to decrease.

[10] https://www.umass.edu/mwwp/resources/factsheets.html

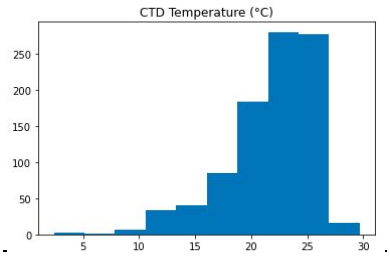
Phosphorus-containing Compounds:

The phosphorus distribution also has a long right tail. Total phosphorus is clustered around the 5-70 ug/L range, while total dissolved has its peak around 3-8 ug/L and soluble reactive phosphorus is around 0.5-1 ug/L. The 2008 Lake Erie Management Plan[11] cited a desired ecological endpoint of <15 ug/L of total phosphorus for the Western Basin of the lake, as an annual average.



Phosphorus compounds seem to follow similar trends to nitrogen compounds over time. They tend to reach lows in late summer, with higher values in spring/early summer and sometimes recovering to high values in fall months. This again could suggest that organisms that fluorish in the summer months are consuming the phosphorus, leading to lower readings at that time.

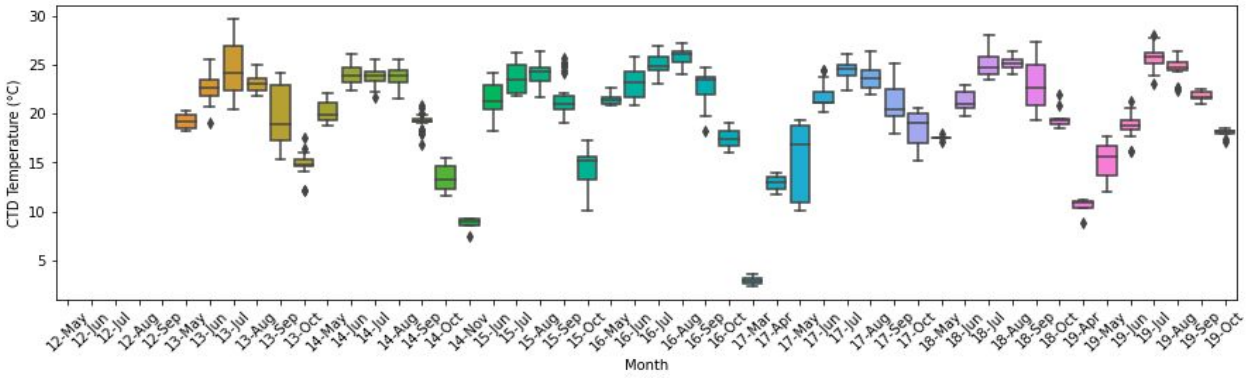
Conductivity-Temperature-Depth (CTD) Measurements:



The temperature distribution is left-skewed which is interesting, but logical due to the fact that most measurements are taken during the summer months, and therefore warmer temperatures. It is also logical that there seems to be a sharp drop above 26°C due to the nature of evaporative cooling. Once a certain temperature is reached (25 to 30°C) it takes a larger energetic input to raise the lake

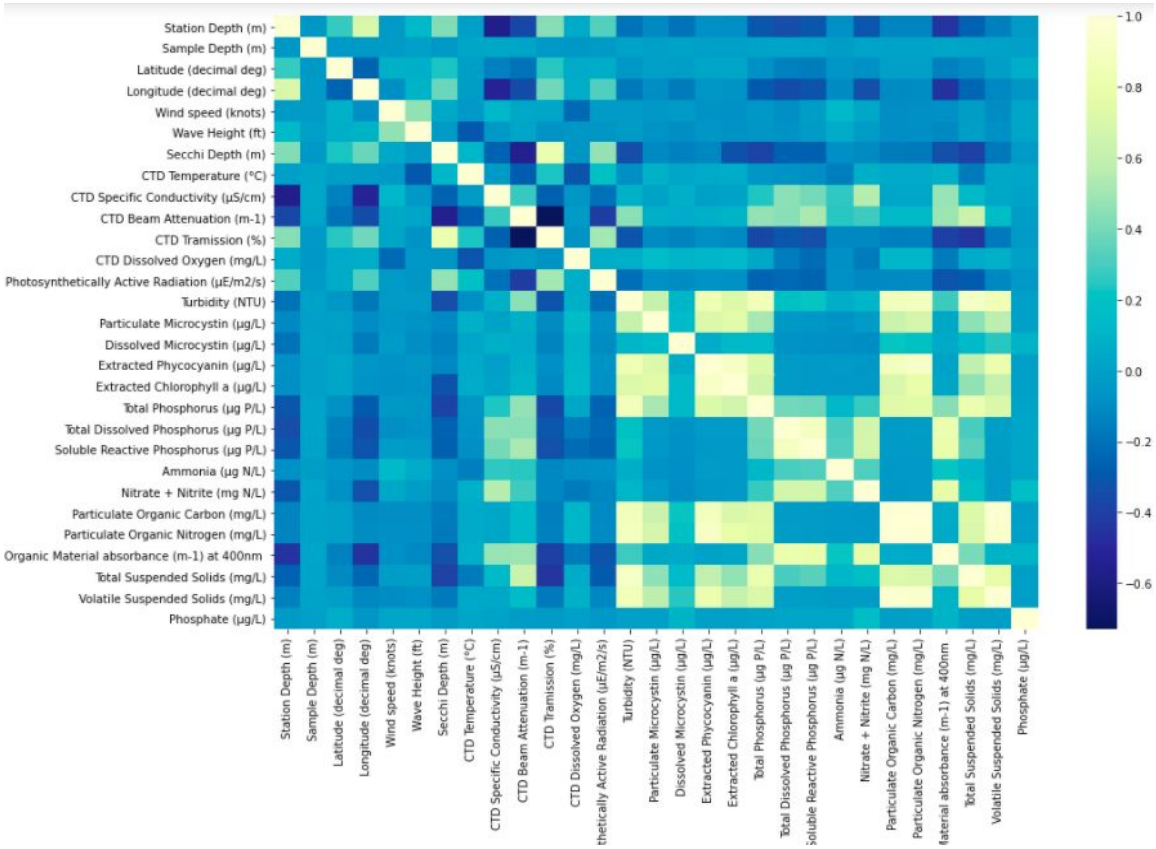
[11]https://www.epa.gov/sites/production/files/2015-10/documents/status-nutrients-lake-erie-basin-2010-42pp.pdf

temperature due to increased evaporation and cooling. Dissolved oxygen was also quite normally distributed compared to chemical nutrients. However, similar to the chemical nutrients, dissolved oxygen tends to start higher in early months, reach a low point in midsummer (July/August) and sometimes recovers in later months (October).



Exploratory Data Analysis - Feature Relationships:

Guiding Question #2: What features are most predictive of an HAB event and how early can one expect to see reasonable indications that an HAB will form?



First a correlation heatmap was visually inspected. Some of the correlations in the heatmap can be explained with straightforward relationships between scientific quantities. For example, transmission and beam attenuation have the defined relationship:

$$C = -(1/r)*\ln(\%Tr/100).$$

Other relationships are not strictly defined mathematically but are expected. For example, Secchi depth is the depth at which a Secchi disk is no longer visible, which should naturally increase with percent light transmission, and decrease with beam attenuation, increasing turbidity, and increasing particulate matter in the water. Turbidity has a large number of strong positive correlations, which also makes sense since turbidity is simply a measure of the cloudiness of the water, which naturally increases as particulate matter increases, and relates directly to transmission and attenuation.

Quantities indicating the growth of algae, phytoplankton, and cyanobacteria such as Phycocyanin, Chlorophyll, and Particulate Microcystin all have positive correlations with each other, as well as weaker correlations with Turbidity and Total Phosphorus. This may at first seem strange since we know that these creatures consume Phosphorus, but Total Phosphorus should include any phosphorus that was incorporated into the actual algae itself. Note that dissolved forms of Phosphorus do not show any correlation with Phycocyanin, Chlorophyll, or Particulate Microcystin.

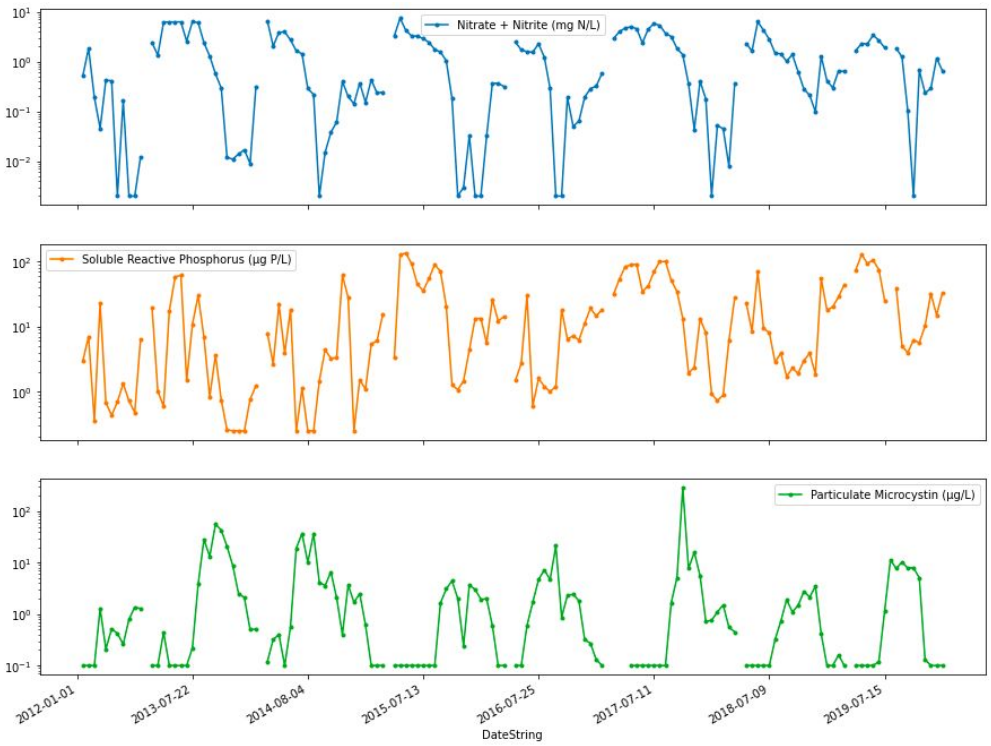
Specific conductivity has some weak positive correlations with various chemical quantities of interest, notably Nitrates and Nitrites, Total Dissolved Phosphorus, and Dissolved Organic Material. This could be due to direct effect, as Nitrates and Nitrites are negatively charged anions, and many dissolved forms of phosphorus and organic material are likely also ions. This could also be due to correlation, since N- and P-containing compounds tend to come from runoff or lake inflows that can also carry other ionic compounds (e.g. halides). Interestingly, besides Total Phosphorus it does not seem like nutrients such as the N-containing compounds or soluble P-containing compounds have any notable correlation with Microcystin amounts. In the case of Total Phosphorus as noted above this may very well be because the Total Phosphorus may include intracellular Phosphorus, and would therefore just be another indication of the presence of an HAB, rather than an independent correlated measurement. The lack of correlation between the Microcystin amount and known HAB nutrients is possibly because of the time lag between the establishment of an HAB-friendly environment (warm temperatures, nutrient-rich) and the growth of the HAB itself. Also, as the HAB grows it consumes nutrients, decreasing their amounts until the bloom can no longer be sustained. This suggests that some sort of time series analysis may be necessary.

Interestingly, Temperature and Dissolved Oxygen don't have many strong correlations, despite their importance to aquatic ecosystems. They have a weak correlation with each other - warmer temperatures correlate with less dissolved oxygen, with the plausible mechanistic hypothesis that warmer temperature promote the growth of aerobic organisms that consume oxygen. Temperature is also naturally correlated with sample depth (deeper samples are colder).

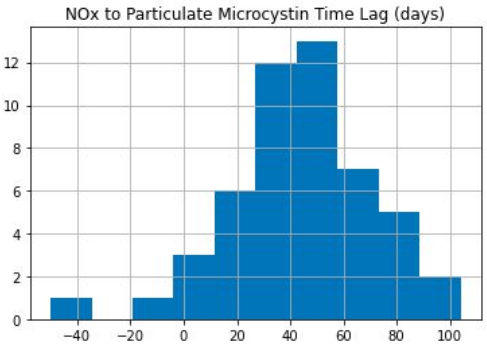
Scatter plots of features showed almost no relationships beyond some nutrient correlations that were noted earlier - soluble forms of phosphorus tend to weakly correlate with nitrates and nitrites. Soluble reactive phosphorus is strongly correlated with total dissolved phosphorus. It was found that there is no clear correlation between the microcystin amounts and the nutrient amounts.

Time Correlations:

Potential time lags between feature values and microcystin response were investigated with stacked line plots. The results for site WE6 are shown here.



There appears to be a consistent offset between nutrient peaks and microcystin peaks. This trend was visually confirmed for several sites near the shore where the highest microcystin concentrations tend to be recorded. At sites further from shore where high microcystin concentrations were less common there was less of a noticeable time lag.

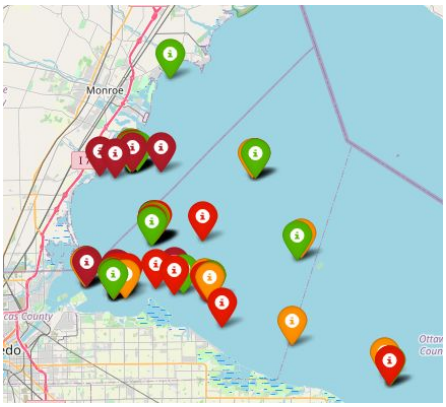


It was difficult to quantify the time lag in an objective way. To keep things simple at this stage we simply looked at the point at which each feature reached its maximum value in a one year cycle for a given location, and calculated at the time lag between those maxima. This was not necessarily the best way to determine time lag. For one thing, local maxima within the year may be

more relevant in some cases than the yearly global maxima. Because nutrient features sometimes recover at the end of a season after the microcystin decreases we could get negative time lags in that case. A more nuanced approach would perhaps fit the data and look at the first derivative of the features in each year, to determine the lag in when nutrient features begin to fall and when particulate microcystin begins to rise. However, as a simple start we were able to obtain estimates that generally ranged from three to eight weeks.

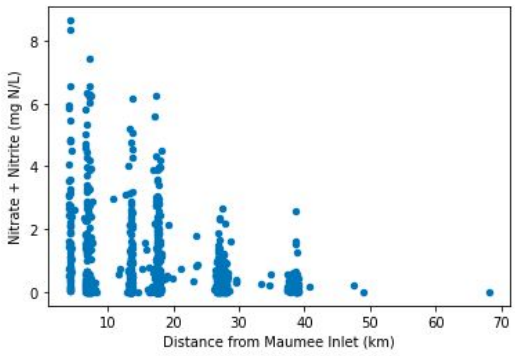
Spatial Correlations:

Geographic influence was explored first through mapping observations and color coding according to particulate microcystin quartile (dark red is the top quartile, green is the lowest). Higher quartile measurements seem to be more frequently located close to shore and close to the Maumee Inlet. This matches intuition on several counts:



1. The water at these locations is shallower and may be expected to reach higher temperatures for longer periods of time.
2. Areas near the shore will have higher concentrations of nutrients due to runoff. In particular the Maumee River that flows into Lake Erie near site WE6 is a major source of nutrient loading into the lake due to agricultural runoff. The river was identified as an area of concern by the EPA 1987 Great Lakes Water Quality Agreement for this reason.[12]

The statistical difference between two sites, WE6 which is close to the Maumee Inlet, and WE13 which is in the center of the western basin, was explored. Clear differences in distributions for microcystin, nutrients, and temperature were found to a 95% confidence interval. The differences between WE6 and WE13 suggested that site matters and should be taken into account, either by grouping sites categorically or using some continuous variable that reflects the differences between sites.



Based on this exploration distances from shore and the Maumee Inlet, a major source of nutrient loading, were calculated for each observation and added as features. The scatterplots here demonstrate that sites that are closer to the inlet have a much higher variance, and higher values tend to be at sites located close to the inlet. This addition allowed us to remove both site labels and coordinates as features, making our analysis more generalizable.

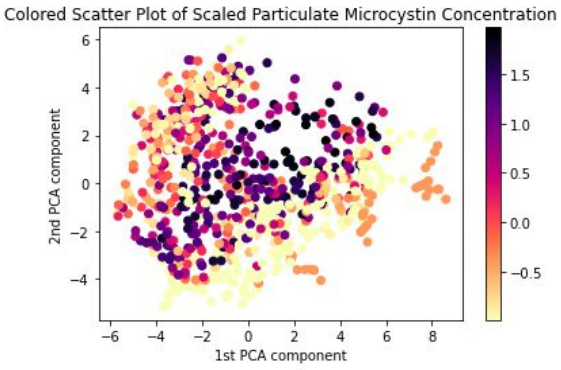
[12] https://19january2017snapshot.epa.gov/maumee-river-aoc_.html

Feature Engineering:

In addition to adding distance features as described above the nitrogen to phosphorus ratio was calculated for each observation. Algae require both phosphorus and nitrogen, and the ratio of nitrogen to phosphorus can be a useful indicator of whether algal growth may be nitrogen-limited or phosphorus-limited. There is some indication that on long time scales for more moderate amounts of algae the algae is typically phosphorus-limited, but during huge bloom events the bloom's overconsumption of nitrogen leads to the bloom becoming nitrogen-limited.[13]

Particulate microcystin had already been identified as the optimal target variable. Features were selected according to their expectation to influence the environment for HAB formation and their independence from indicating HAB formation. The initial selected features were Date, CTD Temperature, CTD Dissolved Oxygen, Total Phosphorus, Total Dissolved Phosphorus, Soluble Reactive Phosphorus, Ammonia, Nitrate and Nitrite, N:P Mass Ratio, Distance from Maumee Inlet, and Shore Distance. Given the probable importance of past nutrient concentrations a time shift was applied to create past features for each observation. This was a small challenge due to the inconsistent frequency of sample collection at different sites. Most sampling was performed weekly or near-weekly, but overall the greatest common factor for intervals between samplings was one day. Therefore the data was filtered by site and year, upsampled to daily frequency with linear interpolation, and time shifts of two, four and six weeks were applied. The data was then resampled back down to weekly frequency using the average as the aggregation method.

As previously discussed most features and the target variable have highly skewed distributions, some over several orders of magnitude. The data were log-transformed and standardized for use in linear regression models. Finally, because the data is ordered in time the test and train split was obtained from unshuffled data, with a test size of 0.2.



Principal component analysis was performed to explore the inherent dimensionality of the data and whether there would be any visible clustering of microcystin levels by the PCA components. About 80% of the data variance is explained by the first five features, and there did seem to be some separation of microcystin levels with just two pca components, as shown here, which suggests that the selected features do have some ability to predict microcystin levels.

[13] https://cen.acs.org/articles/94/i12/Scientists-debate-best-way-tame.html

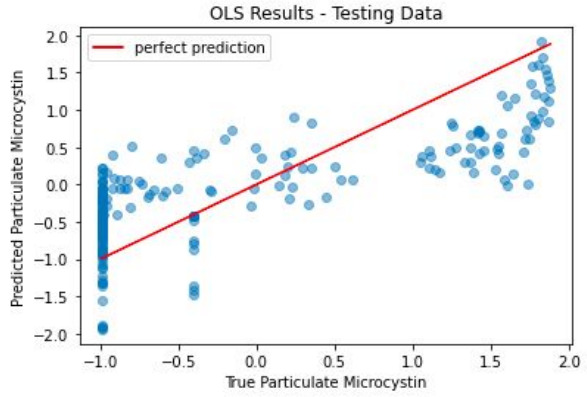
Modeling:

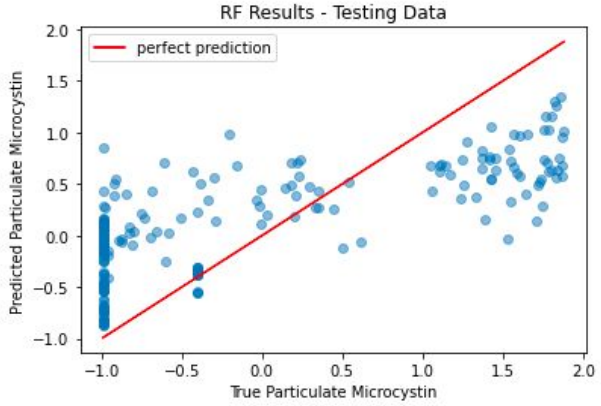
Guiding Question #3: What are the capabilities and limitations of a predictive model for microcystin concentration in Lake Erie?

Regression:

Modeling was first approached as a regression problem to see if the microcystin concentration at a given location on a given day could be predicted, then as a classification problem to see if samples as having low or high microcystin levels could be classified using the Ohio EPA drinking limit of 1.6 µg/L as a cutoff value.

Ordinary Least Squares Regression showed poor performance with a mean absolute error (MAE) of approximately 0.60. The model also showed overestimation of low values and underestimation of high values, as expected due to regression to the mean. It was clear that linear models are far too underpowered for this problem so we moved on to ensemble methods.





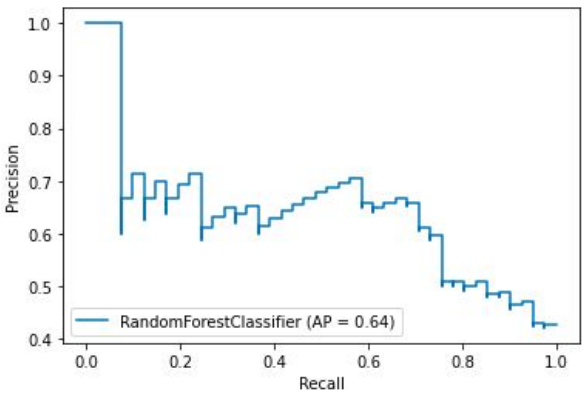
Random Forest and XGBoost regressors were optimized using random hyperparameter search and scaled data, and evaluated on the test set. Both regressors showed very poor performance with MAEs of 0.66 and 0.68, respectively. The MAE here is in log-transformed units since the data was scaled and will correspond to even larger absolute errors on the original data, especially for high microcystin concentrations. Since decision trees use single splits to determine outcomes they don't depend on scaling in principal. However, scaling will affect the optimization of a decision tree for a regression problem since varying transformations will affect the magnitude of the residuals. Given the poor performance of scaled features especially for large microcystin concentrations the unscaled features were explored. The optimized and trained Random Forest and XGBoost regressors again performed poorly on the unscaled data, with MAEs of 1.7µg/L and 1.3µg/L. Neither regressor showed any reliable ability to differentiate between low and high microcystin samples.

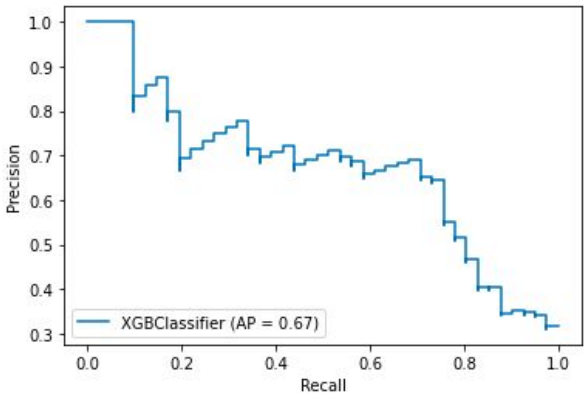
Classification:

While it would have been nice to have a prediction that estimated exact microcystin concentration our goals pertain more to accurate identification and prediction of high concentration samples as a class. From an application and policy standpoint the response to a concentration of 8 µg/L versus 6 µg/L is much the same, what is really important is predicting where a sample will fall in the spectrum of EPA concentration categories. We therefore moved on to binary classification. Although there is not a particular cutoff for particulate microcystin that indicates the definitive presence of a harmful algal bloom we can reasonably use the EPA drinking cutoff of 1.6 µg/L. This cutoff leads to a roughly 20% positivity rate overall. Since the classes are fairly imbalanced hyperparameter tuning was performed with balanced accuracy as the scoring metric.

Classifier Scoring Results		
Classifier	**Precision - Test Set**	**Recall - Test Set**
Random Forest - default threshold	0.79	0.52
XGBoost - default threshold	0.72	0.54
Random Forest - adjusted threshold	0.69	0.71
XGBoost - adjusted threshold	0.65	0.71

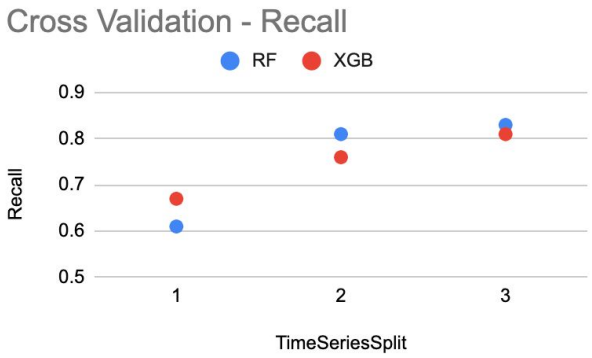
The Random Forest classifier and XGBoost classifier had similar initial recalls of 0.52 and 0.54, but the Random Forest classifier had a somewhat better precision of 0.79 versus 0.72. This is fairly poor performance but both models can distinguish somewhat between very low concentrations and elevated concentrations of particulate microcystin. If the baseline model were to guess randomly using the positive weight as the probability that a sample is a high concentration then we would expect a recall of 0.24, so the models are better than the recall relative to guessing. Recall is the most notable metric for our application since our goal is to identify all HAB events. A false-positive event is an annoyance, but not predicting a bloom that then happens is more detrimental. We explored tuning the model's recall by sacrificing precision through inspection of the precision/recall curves.

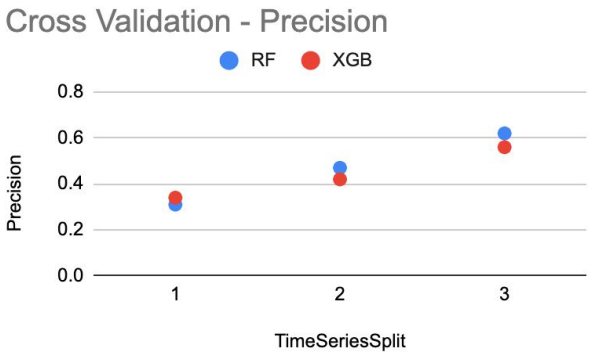




By adjusting the probability threshold of each model we were able to obtain recalls of 0.71 while maintaining precisions of 0.65 for XGBoost and 0.69 for Random Forest on the test set.

Cross-validation with time series splits was performed to get an understanding of the model performance across different training and testing sets. Note that with time series split the size of the training set grows with each successive iteration. The Random Forest classifier initially performs worse than XGBoost on the smallest training set but shows greater improvement as the size of the training set increases, although XGBoost's recall is approaching that of Random Forests in the last set. This suggests that Random Forest may continue to have better improvement than XGBoost for making predictions on future data where the training set would be even larger.





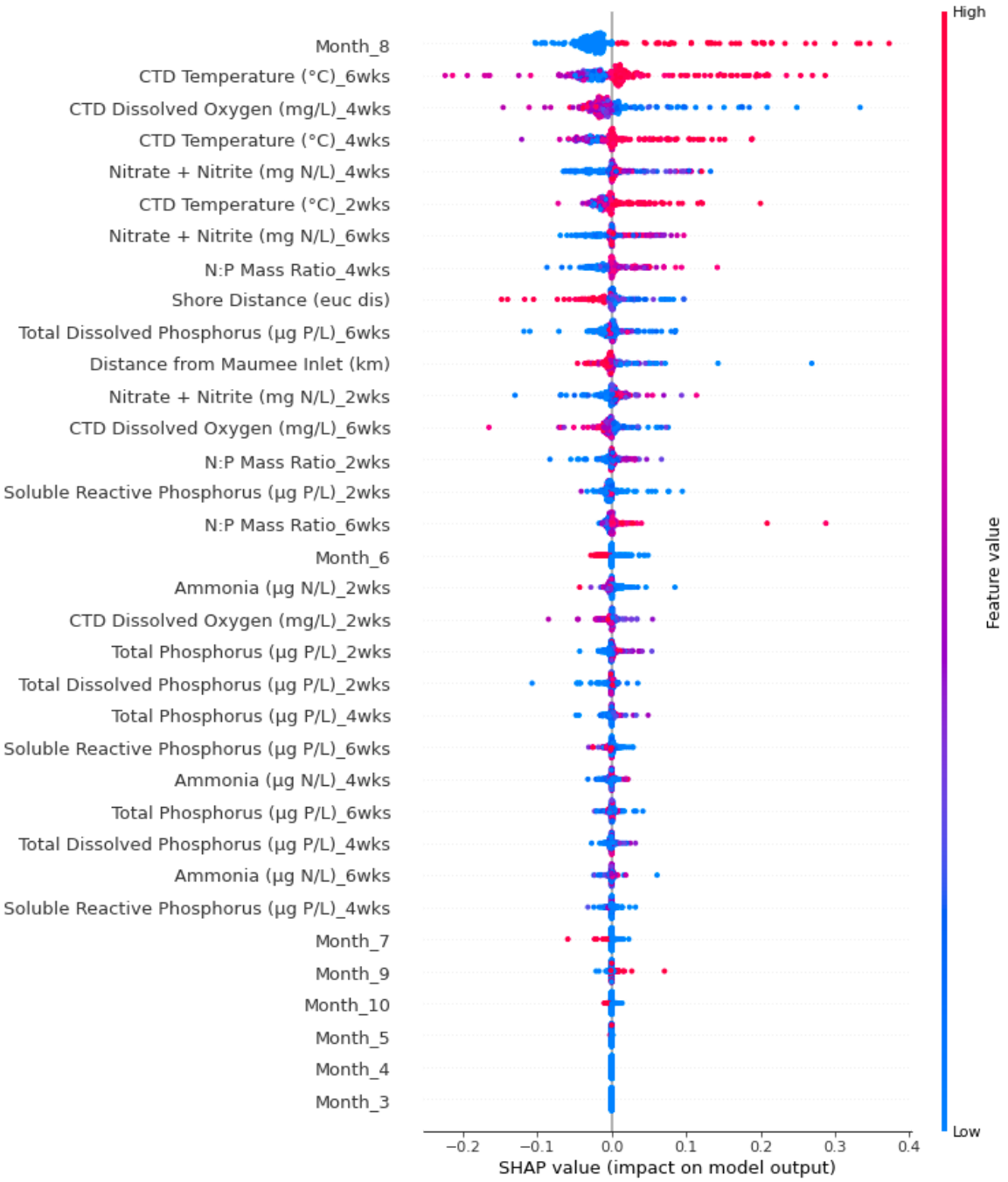
Feature Importance:

Guiding Question #2: What features are most predictive of an HAB event and how early can one expect to see reasonable indications that an HAB will form?

The SHAP package was used to explore feature importance in the Random Forest classifier. For the most part the feature importance is quite intuitive and in alignment with understanding of what factors promote the growth of harmful algal blooms. High temperatures are quite important, and it is especially interesting that high temperatures well in advance of the prediction date are the most influential. It is, however, surprising that there are cases where high temperature lowers the model output. More decisive is whether or not the sample is collected in August, which universally increases model output. Distance from the Maumee inlet and distance from shore both have quite predictable influence - smaller distances increase model output and vice versa, which is logical since closer sites have more concentrated influx of nutrients coming from the river or agricultural runoff.

The presence of nitrogen-containing nutrients increases model output, which is logical since these nutrients promote the growth of HABs. It is interesting that nitrogen-containing compounds were more influential than dissolved phosphorus in this case. It is also quite noteworthy that at more distant times, e.g. 4 and 6 weeks in the past, high nitrogen-containing values increase the model output forecasting HAB presence, but at 2 weeks they are less impactful. This would be in line with the time lag concepts explored in the EDA notebook of nutrients peaking prior to HAB formation, but decreasing as the algae grows because the algae consumes the nutrients. Ammonia in particular follows this pattern, switching from moderate to high values increasing model output at 6 weeks, then high model decreasing model output at 2 weeks.

Finally, high dissolved oxygen decreases the model output consistently. This is interesting because algae are obligate anaerobes, so it would be logical that low oxygen would mean a rather inhospitable environment for a bloom. However, algae also consume oxygen. This is one of the harmful effects that HABs can have - blooms can consume too much oxygen leading to the death of other aerobes. So, it is conversely possible that low oxygen in this case is an indicator of sites that already have eutrophication occurring.



Model Recommendation and Future Work:

Guiding Question #3: What are the capabilities and limitations of a predictive model for microcystin concentration in Lake Erie?

Ultimately for the goals of this project the binary Random Forest classifier with an adjusted probability threshold seems to provide the best recall without sacrificing too much precision based on the test set performance. The XGBoost and RF models showed essentially identical performance but the RF model is slightly better and seems to show greater improvement as the size of the training set increases. Overall this model does provide some ability to predict elevated microcystin concentration based on environmental features. We can expect that ~60% to 70% of sites that the model identifies as having elevated microcystin will actually be contaminated, and that the model will identify ~60 to 70% of such sites. This is somewhat promising performance for a model that does not rely on exact location, or any feature measurements within two weeks of prediction. The model does require on-site measurements for the prediction locations, however. While automated sample collection is currently underway this does limit prediction to a handful of locations. A more useful model would be able to extend prediction to locations between or near by water quality measurement sites. This is one possibility for future work. Another would be to explore the effect of doing away with water quality measurements within the lake itself and instead rely on rainfall, runoff or river loading data. The Heidelberg Tributary Loading Program measures nutrient load into Lake Erie from several rivers of interest, including the Maumee River.[14] The loading data could perhaps be used to develop a simple model for how nutrients diffuse out into the lake and, in conjunction with physical environmental characteristics, eventually cause algal blooms. Such a model would theoretically be applicable to any given location within the Western basin using only data from shore.

[14] https://ncwqr.org/monitoring/