Predicting Blue-Green Algae Quantity in a Semi-Urban Reservoir

Problem Identification

Background

Harmful algae blooms (HABs) occur in lakes across the USA every summer. These HABs, which consist primarily of blue-green algae, are visibly unappealing and can release toxins which can kill fish, waterfowl, livestock, and pets and make humans very ill. Multiple methods exist for quantifying blue-green algae. The most immediate method is to measure phycocyanin, which is a pigment found in cyanobacteria. This sensor is often installed *in situ* and programmed to continuously collect data. Lake operators, such as a state park, will monitor these readings and shut down recreation on a lake if a blue-green bloom is detected. However, these closures are immediate, no warning is available to the public or staff, and all management actions are reactionary. Therefore, lake operators would be very interested in a method to predict blue-green algae in a lake before implementing a shutdown, allowing for planning and allocation of resources.

Objective

Algae, including blue-green algae, can grow to a large volume over the course of days in the right condition. In this project, I developed machine learning models to predict phycocyanin using water quality measurements at an unnamed semi-urban reservoir I named Lake Fictitious. The primary objective was to establish models capable of forecasting phycocyanin multiple days in advance using explanatory variables also collected at the reservoir. With this model and data, the lake operator can predict the magnitude of a HAB before it occurs which would avoid immediate shutdowns.

Data Wrangling

Data Acquisition

Water quality data, including phycocyanin, was collected at Lake Fictitious every 30 minutes for three years using an *in situ* AlgaeTracker manufactured by AquaRealTime. AquaRealTime provided me with a login to their data portal from which I downloaded Lake Fictitious data which consisted of 50,071 rows and 20 columns.

Data Cleaning

The first cleaning step was to remove all columns but those needed for developing a model. These consisted of phycocyanin, sunlight, water temperature, water turbidity, and chlorophyll-a.

Next, I had to address the many rows with the same date and time. Rather than simply deleting the second occurrence, I calculated the mean of the columns in the two rows because column values for multiple duplicate pairs were slightly different. During this process, I discovered that the data was collected every hour at the beginning of the study but switched to every half and hour at some point.

No data was missing from existing rows in the dataset. However, many rows were completely missing. I inserted these rows with the appropriate date and time but left teh values as NaNs for the time being. The largest of these data gaps was 6.25 days which is not too large to impute data which I performed in later steps.

I calculated descriptive statistics for the metrics (Table 1). The values for sunlight were realistic while the maximum values for the other metrics are possible but very unlikely. These readings are likely the result of some interference (i.e. debris) on the sensors.

Table 1. Metric descriptive statistics.

Stat	Phycocyanin (RFU)	Sunlight (PPF)	Water Temperature (°C)	Turbidity (NTU)	Chlorophyll-a (RFU)
Mean	117	165	18.7	62.5	635
STD	173	2,473	6.3	101	703
Min	0	0	0.0	0	0
Max	1,655	2,442	43	400	3,000
Median	55	<1	18	17	308

To investigate these max values and look for other potentially erroneous readings, I plotted the data with markers for each datapoint and lines connecting continuous data (Figure 1). I made the following observations in the plots:

- Readings were collected every hour until roughly 6/10/21 when collection switched to every half hour.
- The hourly readings at the beginning of the dataset for sunlight are all very low. These values would be expected at night but continuously for four months.

- Around 12/1/21, the water all metric values are very extreme. AquaRealTime informed me that the unit was flipped upside down and the recorded values are incorrect.
- Water temperature values in 4/23 are not reliable at zero degrees.
- Other maximum readings for phycocyanin, water temperature, turbidity, and chlorophyll-a are also not reliable because these values are unlikely.

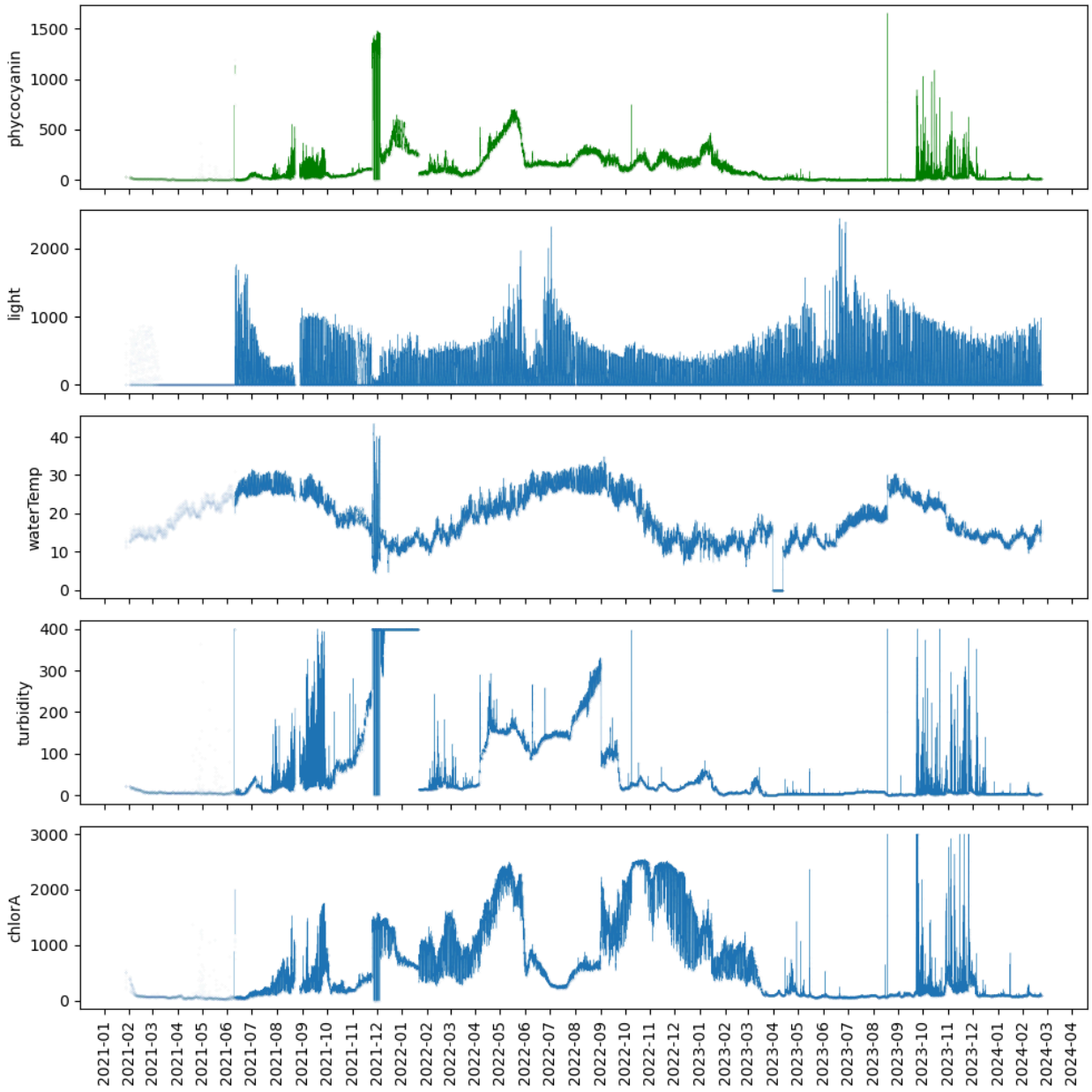

Figure 1. Plots of uncorrected metric data.

I addressed these unreliable data through many steps. First, I replaced the low hourly increment sunlight data, the values when the unit was flipped, and the low water temperature data with NaNs. I also replaced the unrealistic maximum values for each

metric with NaNs. For sunlight, I replaced these NaNs and a gap in the data at the end of 8/21 with the mean values of the corresponding dates in 2022 and 2023. Linear interpolation would not be appropriate for this big data gap because light is high during the day and at zero every night and a straight line would not capture this day/night complexity over the time period. I then used spline linear interpolation to fill the data gaps I created in the other metrics and for all other NaNs in all metrics. This interpolation tries to ensure smoothness and continuity in the interpolated curve. In the end, my dataset contained no NaNs.

Exploratory Data Analysis

Visualization

I plotted the newly corrected metric data to assess the metrics for patterns and trends (Figure 2). As would be expected, sunlight and water temperature values follow a seasonal pattern of being higher in the summer and lower in the winter. June was cloudy every year and 2021 was also cloudy in July, Aug, and Dec.

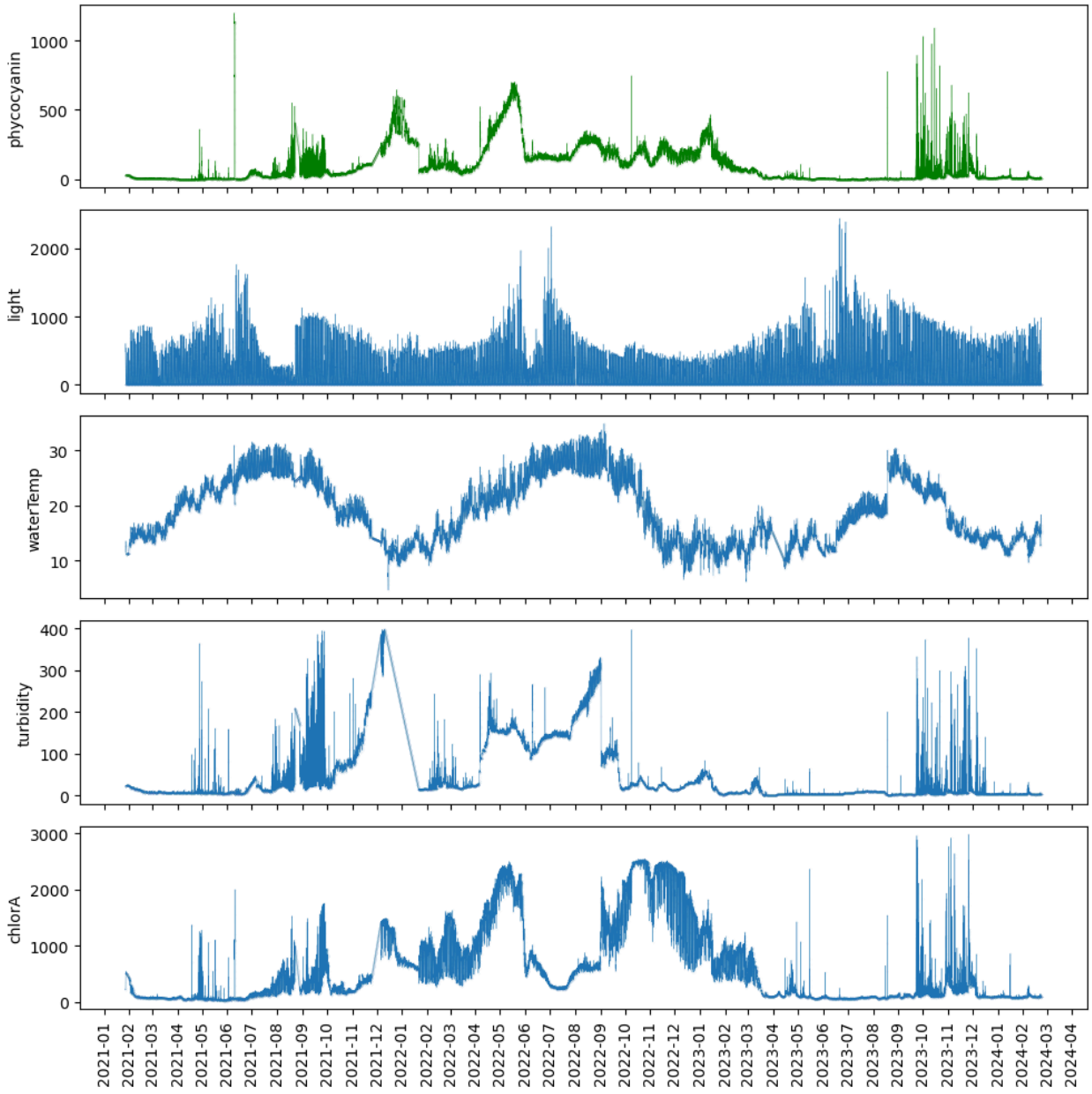

Figure 2. Plots of corrected metric data.

Phycocyanin, turbidity, and chlorophyll-a did not appear to be seasonal and, in fact, the three years were very different from each other. Values often increased for short periods (days) but generally returned to baseline or close to baseline. Exceptions to returning to baseline occurred for turbidity from roughly 10/21-2/22 and 4/22-10/22, for phycocyanin and chlorophyll-a from roughly 12/21-2/22 and 4/22-6/22, and for just chlorophyll-a from roughly 9/22-1/23. Overall, baseline was higher from 11/21-2/23 for all three metrics. This variation from year to year is likely the result of 2021 being a cloudier year than 2022 and 2023 and from rainfall and nutrient inputs which were not a part of this study.

I also created a boxplot for each metric (Figure 3). All values for water temperature fell within the whiskers (1.5 times the IQR). Many readings for phycocyanin, turbidity, and chlorophyll-a were greater than three standard deviations from the median which is a common way to define outliers. As erroneous data has already been removed, there's no reason to believe that any of these data should be considered outliers and removed from the dataset.

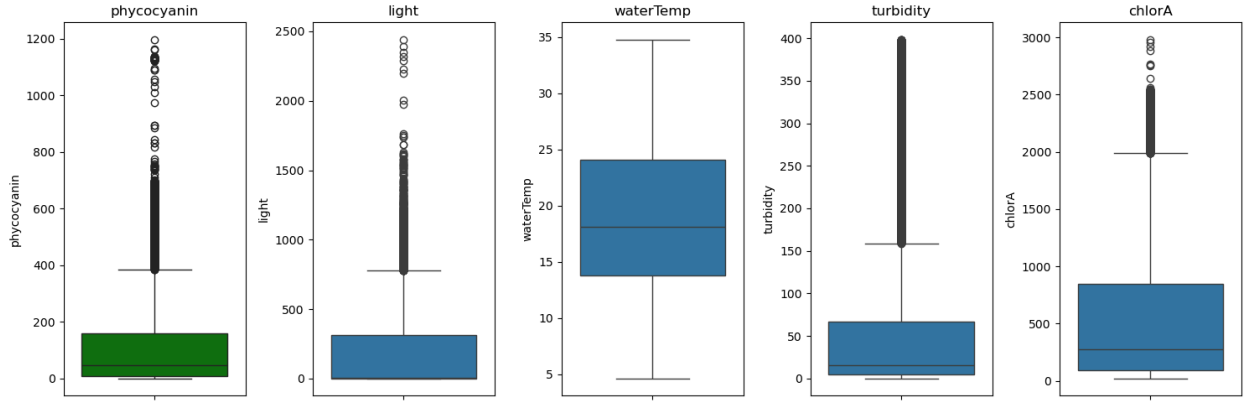

Figure 3. Boxplots of corrected metric data.

Lastly, I created histograms for each metric (Figure 4). These plots largely reflected the boxplots. No metric appeared to be normally distributed, however, water temperature was closer than the others. All other metrics were right-skewed.

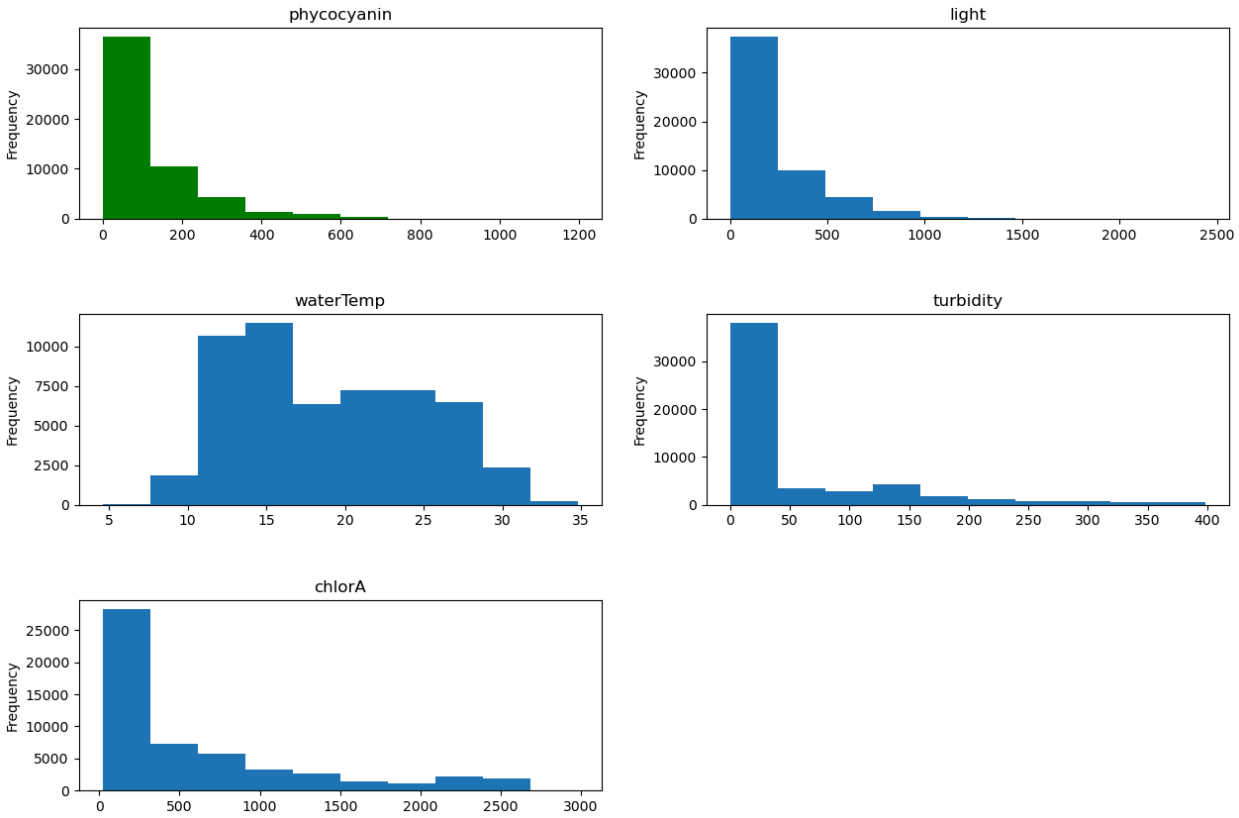

Figure 4. Histograms of corrected metric data.

Normality

Shapiro-Wilk test results for each metric had p values of 0.0 indicating not normal distributions. This agrees with the histograms.

Correlations

I created a correlation heatmap between metrics to investigate linear relationships, specifically with phycocyanin (Figure 5). Heatmaps are an effective way to visualize many correlations at once, making it easy to identify correlations and patterns in correlations. The heatmap showed some correlations between the metrics. Moderately strong correlations occurred between phycocyanin and turbidity and between phycocyanin and chlorophyll-a. Moderately weak correlations occurred between turbidity and chlorophyll-a. The metrics correlated with phycocyanin are likely to be good predictors in the models. Conversely, the metrics with a low correlation to phycocyanin are likely to be less relevant in the model and could be excluded if needed during model development. No metric was very correlated (close to 1) with phycocyanin, which would have warranted removal from the data. I also created a pairplot (Figure 6)

to further investigate linear relationships and also other relationships between phycocyanin and other metrics but no new observations were made.

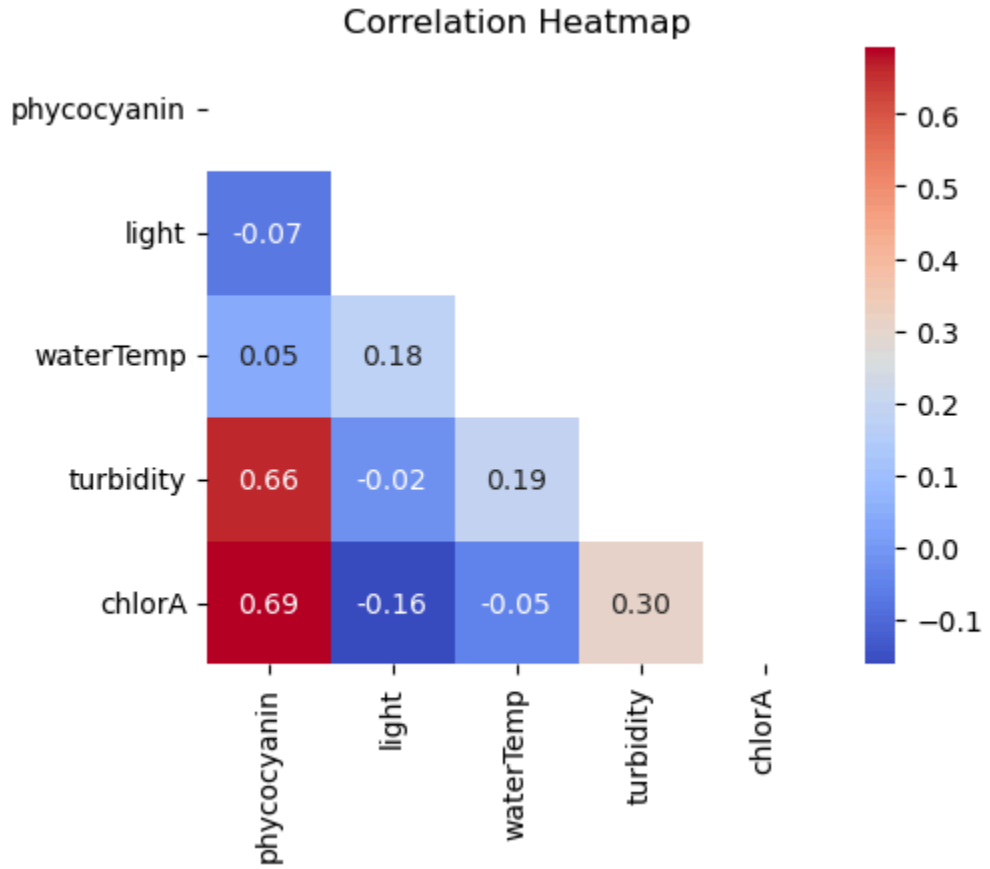

Figure 5. Correlation heat map of metric data.

Figure 6. Pairplot between SVI and other metrics.

Stationarity

To investigate stationarity, I began by detrending phycocyanin (Figure 7). The trend plot showed no long-term progression or direction of the data over time. In addition, the seasonal plot showed a daily fluctuation pattern which occurred in a regular daily pattern. Lastly, the residual plot shows the random noise left in the data after removing the trend and seasonal components. Unexplained variability occurs periodically throughout the study period and is most intense in 10/23-1/24.

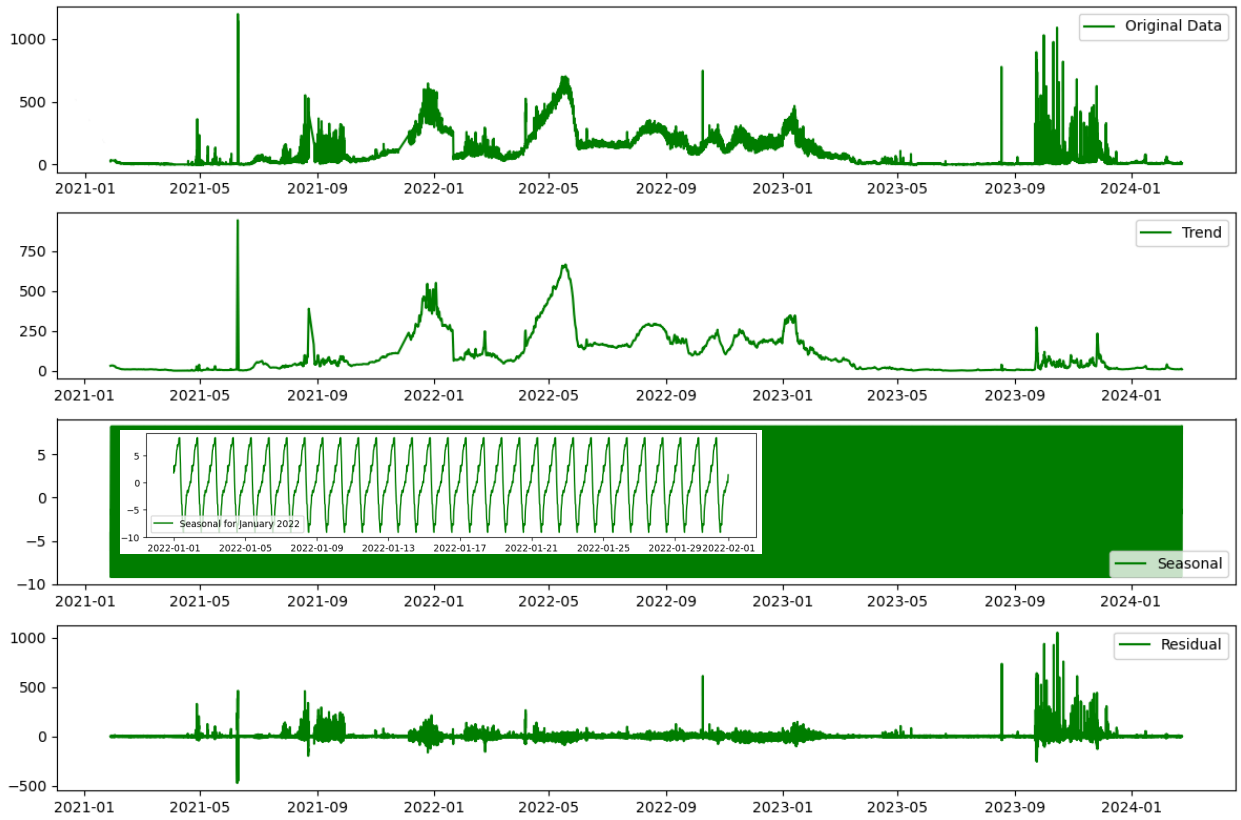

Figure 7. Decomposition plots of phycocyanin. The seasonal trends appear as a block because they are very tight. The inset graph of 1/22 provides more resolution to see the daily trends.

Rather than print the detrending graphs for each metric, I ran KPSS and ADF tests for each metric to test for stationarity. The p-values for all metrics in the KPSS tests were <0.01 which is less than the significance level of 0.05. Therefore, I rejected the null hypothesis of stationarity for all metrics meaning they are all likely non-stationary.

The ADF test p-values for water temperature, turbidity, and chlorophyll-a metrics were greater than 0.05. Therefore, I failed to reject the null hypothesis. This suggests that these metrics may have a unit root and could be non-stationary. However, the p-values were all less than 0.05 for the phycocyanin and sunlight metrics, indicating strong evidence against the presence of a unit root. I reject the null hypothesis for these metrics, suggesting that they are likely stationary according to the ADF test.

Together, the KPSS and ADF tests indicated that the temperature, turbidity, and chlorophyll-a metrics are non-stationary. As for phycocyanin and sunlight, the KPSS test indicated that these metrics are non-stationary while the ADF test indicated that they are stationary. This apparent contradiction indicates that these three metrics are difference stationary. As a result, all metrics will need to be detrended through differencing techniques.

Pre-processing and Training Data Development

Grouping Rows

Typically, time series data granularity, or unit, should be the same as the unit of the desired prediction, in this case, days. Therefore, I calculated the mean values of the metrics over a one day period so I could use these data for my models. This step can also sometimes normalize and/or make non-stationary data into stationary data. As a result of this process, the data frame was reduced to 1,123 rows.

Feature Engineering

Normalize

I reran the Shapiro-Wilks test results for each metric. P-values were just above zero which was higher than previously but still far below 0.05 indicating not normal distributions. Many machine learning models require input data to be normalized.

The natural logarithm is often used for normalizing data, especially when the data has a heavily right-skewed distribution. Taking the natural logarithm can compress large values, making the distribution more symmetric and closer to a normal distribution. In addition, the natural logarithm is often used for detrending stationary data.

After performing natural logarithms, I reran the Shapiro-Wilk tests and all p-values were greater than 0.05 indicating normally distributed data.

I again ran the KPSS and ADF tests on the grouped and normalized data and found the results to be similar to before. The only difference was that both tests found the sunlight data to be stationary.

Temporal

All metrics but sunlight required detrending so I performed differencing to compute the differences between consecutive data points. This technique is often performed for time series models and can transform a non-stationary time series into a stationary one and remove trends and seasonality. Overall, differencing can be beneficial because it stabilizes the mean and variance of the data, making it easier to identify underlying patterns and relationships.

I then ran the KPSS and ADF tests on the new features. The results indicated that all differenced features were stationary.

Lag Features

Creating lagged versions of data is also a common technique performed for time series models and incorporates temporal relationships involving past values into the model. I created 1 through 20 day lagged versions of the original features and the transformed (natural log and differenced) features. Lagging was done simply by copying and shifting the columns in the dataframe the appropriate number of days.

Pre-processing

I defined two different sets of X and y. One includes just the original and lagged features of the original data. The other, includes just the transformed features and lagged features of the transformed data. These two sets allowed me to compare the performance of the models using unaltered and transformed data. I chose to leave the the lagged phycocyanin features in X for the models because autocorrelation may be large part of forecasting phycocyanin.

Modeling

Models capable of making predictions from time series data were essential for this project. I chose to explore linear regression, random forest, and XGBoost models.

For each model, rather than splitting, training, and testing once, cross-validation was performed to divide the data into 5 subsets and training and testing were performed five times using a different subset for testing each time. I specifically used TimeSeriesSplit() for my cross-validation because this function is specifically designed for time series data. Each fold is a superset of the previous one, ensuring that the model is trained on data that retains the temporal nature of time series data. Each fold was scored and averaged for a final score. This process avoids overfitting the model to the data and thus makes it more generalized for future data. The score I used was root mean squared error (RMSE) which measures the error between the model's predictions and the actual values.

Each model was run multiple times starting with including non lagged feature data and adding a level of lagged feature data with each iteration up through the 20 lag features. This process was performed for both sets of X and y. The RMSEs for each iteration were graphed and I was able to identify the number of lags for the model that would produce the lowest RMSE. The lowest RMSE from the models using the transformed data had to be *un*-transformed (*un*-natural logged and *un*-differenced) before comparing to the other RMSEs so that they were on the same scale. Once scaled correctly, I could then compare these best RMSEs between models.

Multivariate Linear Regression

First I ran the linear regression models using the original data and an increasing number of lagged days (Figure 8). The data indicates that using the original data and the data lagged 1-3 days produces the best linear regression model. A score of 29.5 is approximately 1.8% of the phycocyanin range (median was 55, mean was 166.73, min was 0.00, and max was 1654.64) which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 29.5 RFU would not constitute a blue green bloom at a dangerous level.

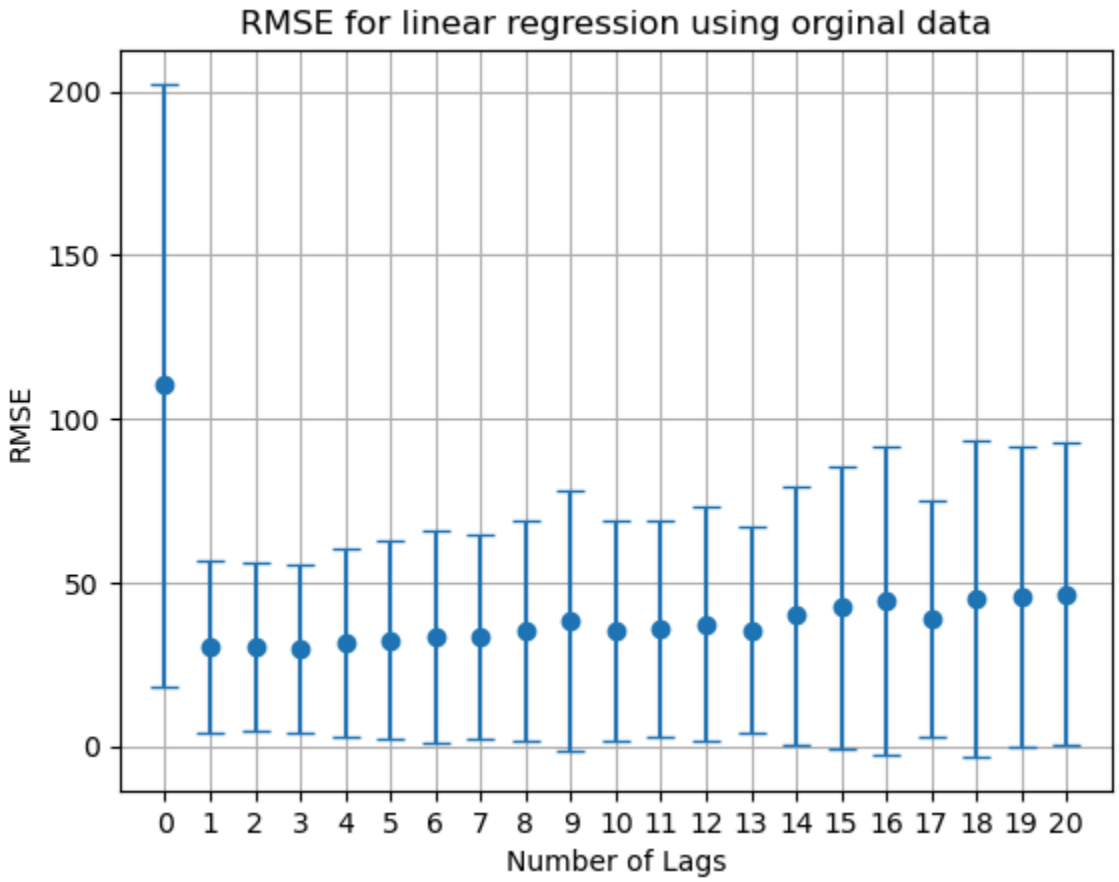

Figure 7. Root mean squared error for each linear regression model run with the original data and increasing number of lags. Whiskers are for standard deviation.

Next I ran the linear regression model using the transformed data and an increasing number of lagged days (Figure 8). The data indicated that using the transformed data with no lagging produces the best linear regression model. After *un*-transforming the RMSE, I could compare it to other RMSEs. A score of 14.4 is approximately 0.9% of the phycocyanin range which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 14.4 RFU would not constitute a blue green bloom at a dangerous level.

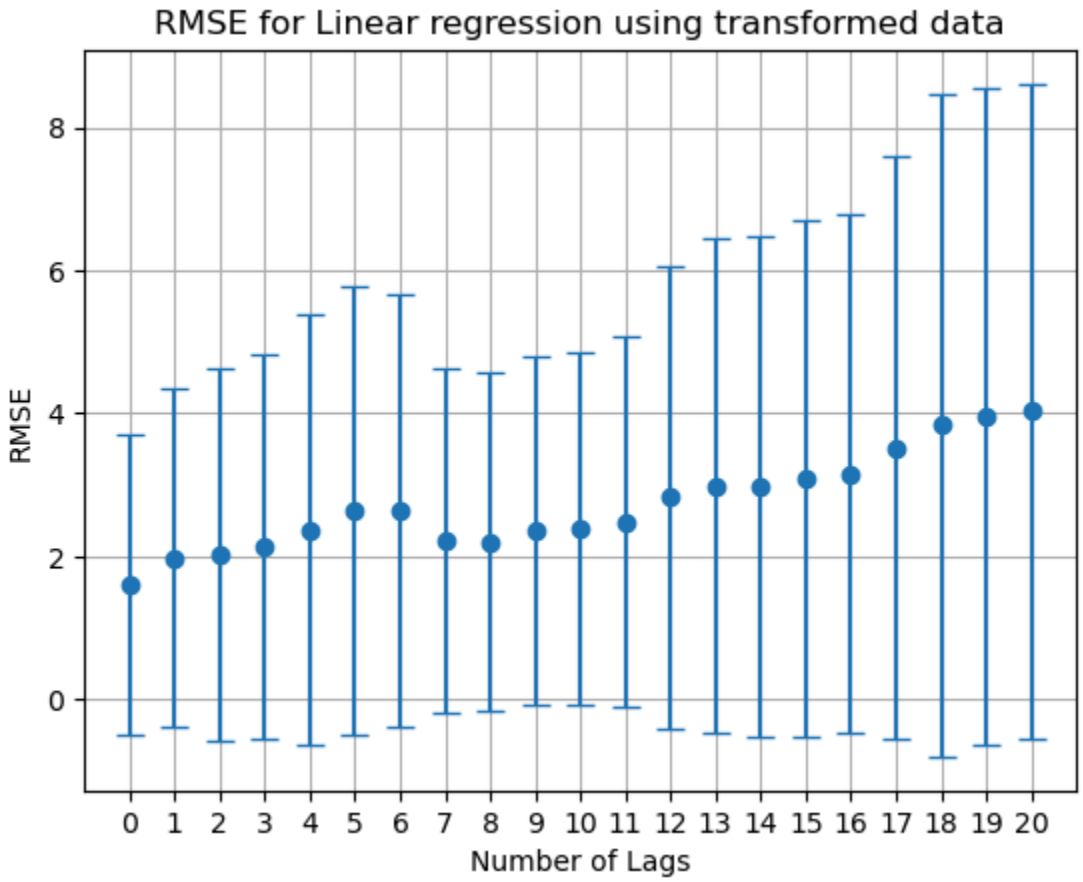

Figure 8. Root mean squared error for each linear regression model run with the transformed data and increasing number of lags. Whiskers are for standard deviation.

Random Forest

First I ran the random forest models using the original data and an increasing number of lagged days (Figure 9). The data indicated that using the original data and the data lagged 1-13 days produced the best random forest model. A score of 43.7 is approximately 2.6% of the phycocyanin range which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 43.7 RFU would not constitute a blue green bloom at a dangerous level.

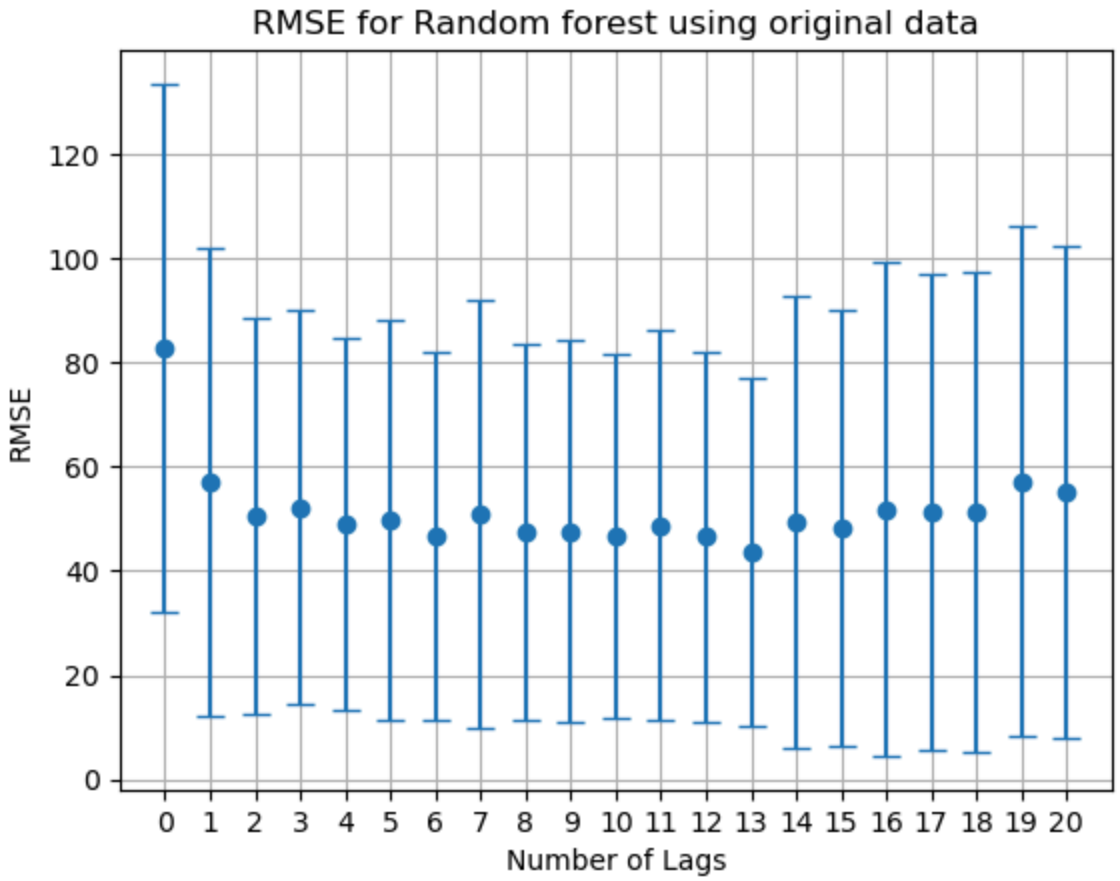

Figure 9. Root mean squared error for each random forest model run with the original data and increasing number of lags. Whiskers are for standard deviation.

Next I ran the random forest models using the transformed data and an increasing number of lagged days (Figure 10). The data indicates that using the transformed data with 0 to 8 day lagged features produces the best random forest model. After *un*-transforming the RMSE, I could compare it to other RMSEs. A score of 12.8 is approximately 0.8% of the phycocyanin range which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 12.8 RFU would not constitute a blue green bloom at a dangerous level.

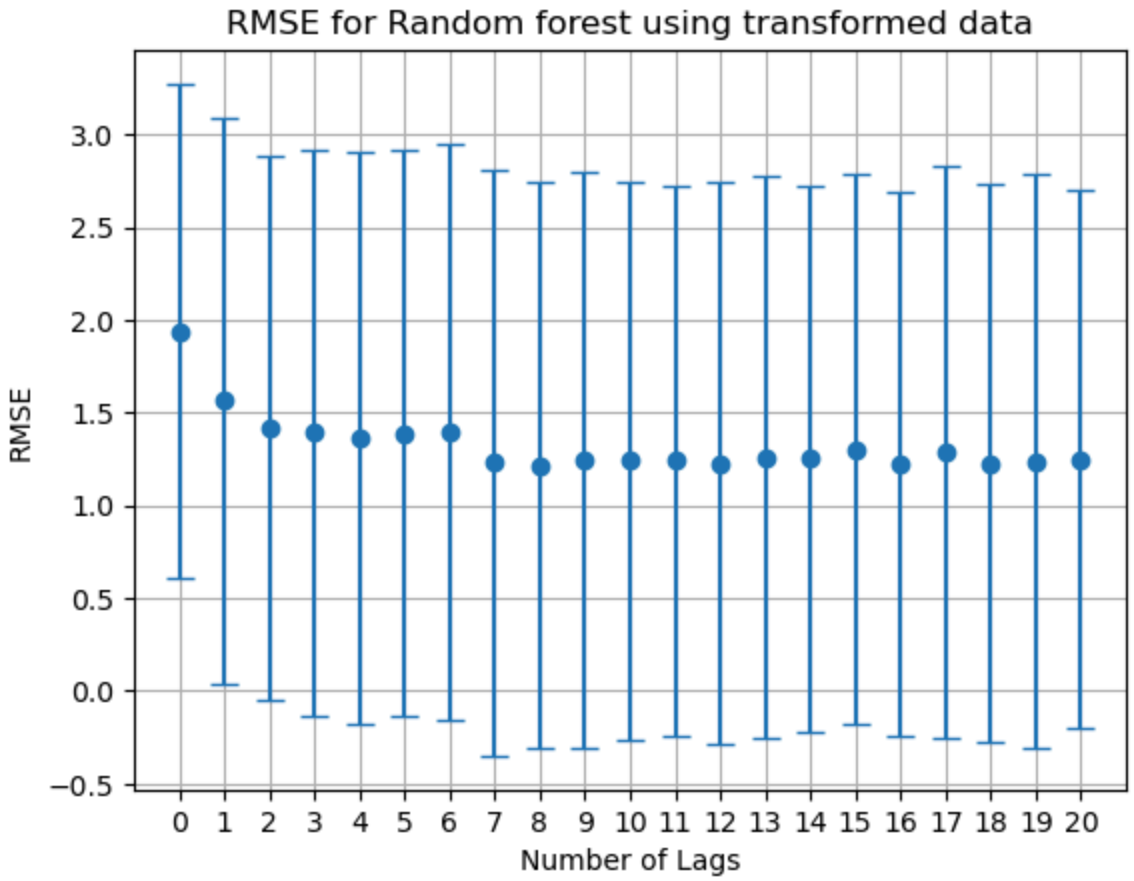

Figure 10. Root mean squared error for each random forest model run with the transformed data and increasing number of lags. Whiskers are for standard deviation.

XGBoost

First I ran the XGBoost models using the original data and an increasing number of lagged days (Figure 11). The data indicates that using the original data and the data lagged 1-14 days produces the best XGBoost model. Adding more lagged data improves the RMSE slightly but not enough to justify making the dataset more dimensional. A score of 55.3 is approximately 3.3% of the phycocyanin range which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 55.3 RFU would not constitute a blue green bloom at a dangerous level.

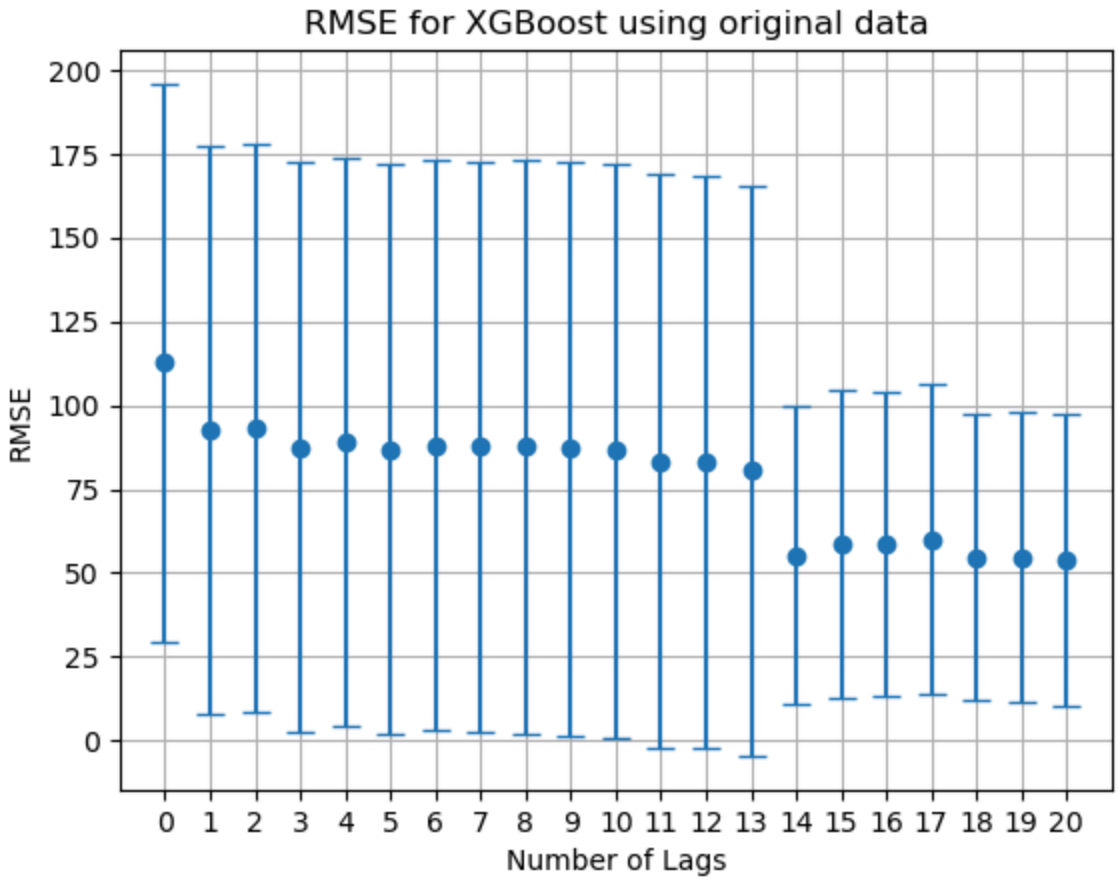

Figure 11. Root mean squared error for each XGBoost model run with the original data and increasing number of lags. Whiskers are for standard deviation.

Next I ran XGBoost models using the transformed data and an increasing number of lagged days (Figure 12). The data indicates that using the transformed data with 1-10 day lagged features produces the best XGBoost model. After *un*-transforming the RMSE, I could compare it to other RMSEs. A score of 12.5 is approximately 0.8% of the phycocyanin range which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 12.5 RFU would not constitute a blue green bloom at a dangerous level. Overall, this model performed better than all other models.

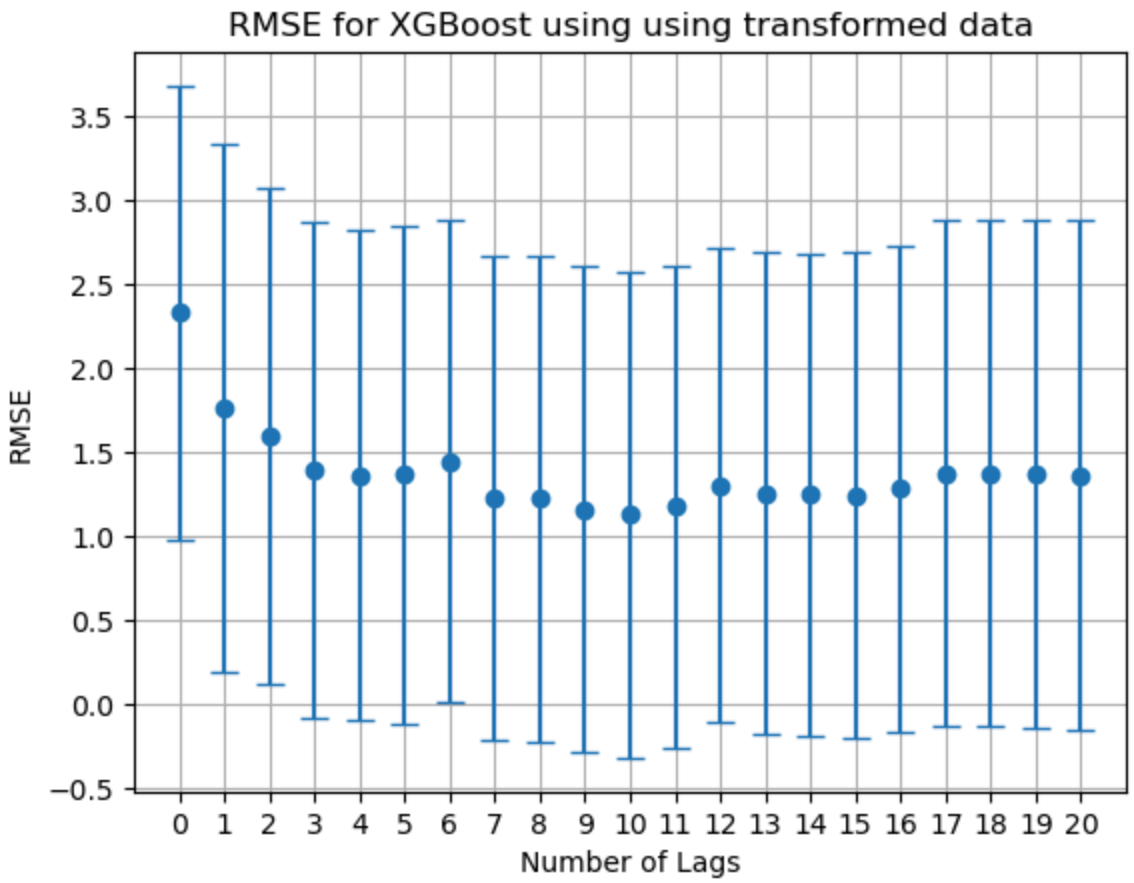

Figure 12. Root mean squared error for each XGBoost model run with the transformed data and increasing number of lags. Whiskers are for standard deviation.

I then performed hypertuning on this best model. I did this using a random search on many of the hyperparemeters for the XGBoost model. The model improved slightly to 12.4. This score is approximately 0.7% of the phycocyanin range which is considered low error. In other words, the error is not worrisome because the median of 55 +/- 12.4 RFU would not constitute a blue green bloom at a dangerous level.

Conclusion

Overall, the tuned XGBoost model using transformed data lagged at zero to 11 days produced the highest performing model to predict phycocyanin in Lake Fictitious. While error in the prediction existed, it is small in the grand scheme of HABs where phycocyanin values are, at a minimum, in the several hundreds. A false positive of a lake operator closing a lake when phycocyanin is forecasted to be 200 RFU but, in fact, due to the error in the model, the real value is 187.6 RFU, is acceptable. Therefore, this model could be very helpful for predicting blue-green algae in a lake before implementing a shutdown, allowing for planning and allocation of resources.