Water Quality in the Lower James Can Help Explain Phytoplankton Levels in the Chesapeake Bay

An Analysis by Matt Carswell

12/15/2023

Chapter 1: Introduction

According to a recent study published by Stanford University and the National Autonomous University of Mexico, we are currently experiencing a sixth mass extinction event. [1] A mass extinction event is defined as the planet losing 75% of its species within a "short geological time period." We are currently seeing species of all kinds, plant and animal alike, becoming extinct at a rate unprecedented to the planet since the dawn of the human race. This extinction event is being driven by one major entity: anthropogenic climate change.

Upon reading this study, I became curious as to how this sixth mass extinction event might be impacting me individually, specifically, how it might be impacting the areas around my hometown of Mechanicsville, Virginia, a suburb of the greater Richmond area. I decided to take a look at how phytoplankton, an important life force of the Chesapeake Bay Watershed, is currently being affected by water quality in my own local watershed, the Lower James Watershed.

To review specifically what this project will explore, per the presentation I delivered for this class, I will be using two datasets from the Chesapeake Bay Data Hub, one that contains water quality data from a monitoring station close to my hometown and another that contains data on phytoplankton counts from both the same Lower James test monitoring station and also a monitoring station in the Lower Chesapeake Bay watershed. [2] The main questions of interest include:

- Can we identify and quantify the impact of climate change/anthropogenic pollutants on the Lower James Watershed?
- What variables explain variation in phytoplankton levels the most?
- Can we link patterns of water quality in the Lower James to effects in the Chesapeake?

Chapter 2: The State and Drivers of Water Quality and Plankton in The Lower James and Chesapeake Bay

2.1 Historical Water Quality of the Lower James

This investigation started by looking at a variety of measures of water status and quality recorded from the chosen monitoring station (TF5.2A) from December 1985 to March 2023. The measures of interest chosen for this investigation including depth, dissolved oxygen (DO), pH, total nitrogen (TN), total suspended solids (TSS), and water temperature (WTEMP).[3] We can observe theses measures over the specified date range in **Figure 1**.

[1] https://news.stanford.edu/2023/09/18/human-driven-mass-extinction-eliminating-entire-genera/
[2] https://datahub.chesapeakebay.net/Home
[3] Some rationale for picking these specific measures can be attributed to the following link: https://sinay.ai/en/what-are-the-main-indicators-of-water-quality/

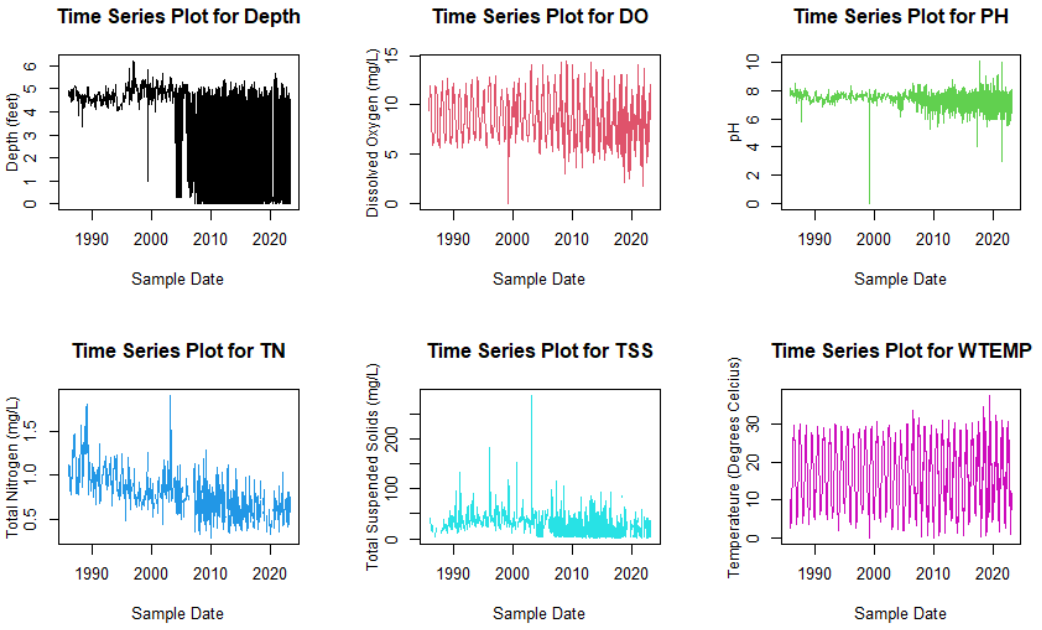

Figure 1: Time Series Plots of Lower James Water Features

We can see that our plots contain a lot of noise (which is to be expected), but we can pick out that there might be a decrease in pH, TSS, and DO in the Lower James over time, and definitely a decrease in TN. In order to more clearly observe these trends, a 5-year moving average was created for each of these variables and then plotted in **Figure 2**.

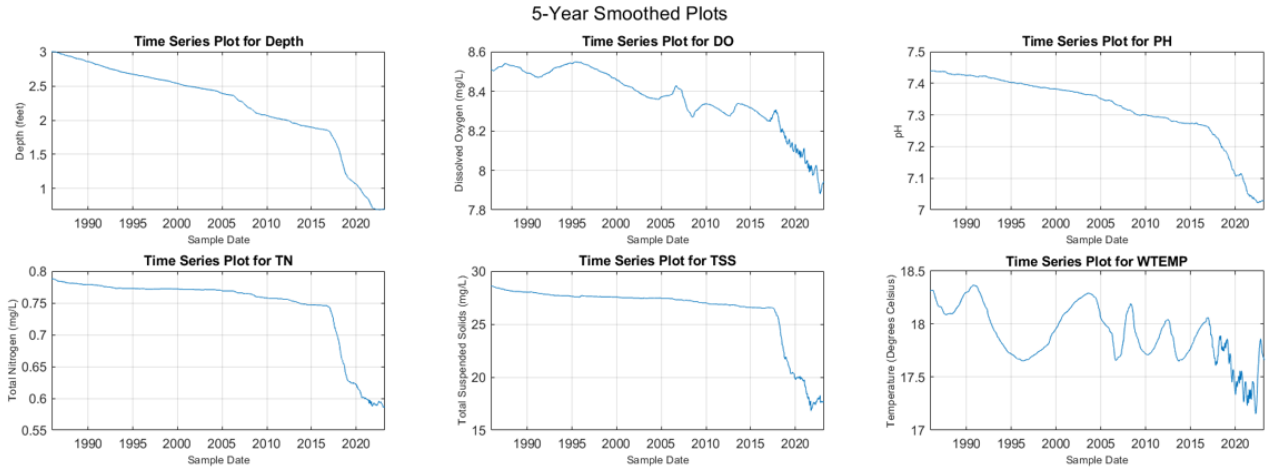

Figure 2: 5-Year Smoothed Plots of Lower James Water Features

It can now be observed that there is a definitive decrease in ALL of our features over time. Also, each of them seem to present a relatively steady decrease up until around the middle of the 2010s where there then is a more significant decrease, which is obviously a point of curiosity. We are going to be looking more at how these variables are specifically correlated with each other later in this report, but early exploratory analysis tells us the Lower James is currently resting at historically low levels of

depth and DO. This portion of the river is also coming right off of historically low water temperatures in 2020. Low water depth and DO levels create worse conditions for life in the water to thrive,[4] allowing us to hypothesize that perhaps these phenomena are negatively impacting the Lower James and Chesapeake ecosystems, especially phytoplankton. Also, while lower water temperatures are not usually associated with climate change, drastic changes in temperature, either direction in magnitude, can still be associated with the effects of climate change, destabilizing the conditions in which native species are used to subsisting upon.[5]

There are some positive takeaways from these graphs, however, in that TN and TSS have begun to approach even safer levels over the years along with the pH of the water becoming more neutral. These observations are signs that environmental protection policies in Virginia, both at the state and municipal levels, are working and anthropogenic pollution is being seriously mitigated (at least in central and southeastern Virginia). Low TN levels, however, might still be detrimental to the ecosystem which will be touched on in later sections.

2.2 Phytoplankton in the Lower James and Lower Chesapeake Bay

The second component of this investigation, as mentioned, involves looking at phytoplankton counts in both the Lower James and also Lower Chesapeake. In order for a comprehensive look at phytoplankton in these watersheds, both the total count of individual phytoplankton AND the amount of different phytoplankton species in the water, a measure of biodiversity, were analyzed. The time series data for these two variables, recorded from March 1986 to December 2021, were plotted below in **Figure 3.**

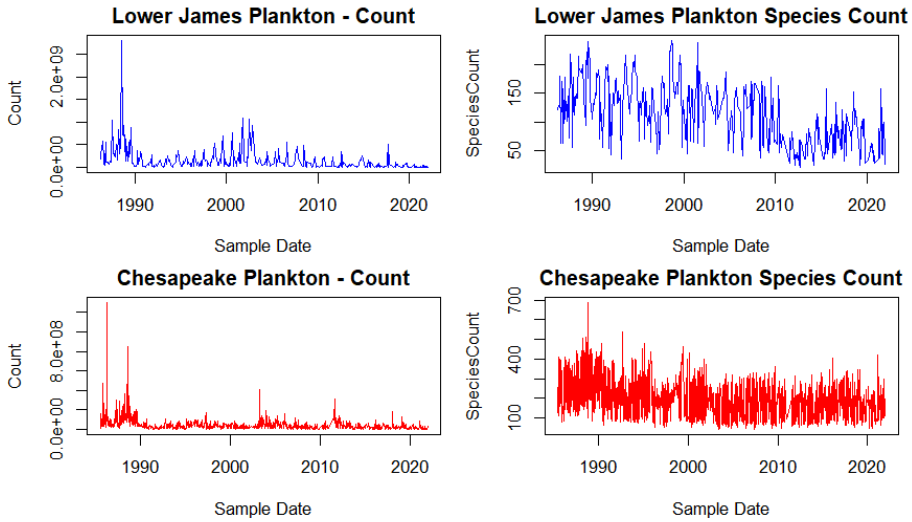

Figure 3: Time Series Plots of Phytoplankton Counts

[4] Relating to impact water depth: https://homework.study.com/explanation/how-does-water-depth-affect-marine-life.html#:~:text=The%20water%20depth%20affects%20marine,them%20and%20damage%20their%20bodies. DO: https://sarasota.wateratlas.usf.edu/library/learn-more/learnmore.aspx?toolsection=lm_dissolvedox#:~:text=It%20is%20essential%20for%20the,other%20aquatic%20organisms%20cannot%20survive.

[5] https://royalsociety.org/topics-policy/projects/climate-change-evidence-causes/question-11/

These plots show us that there might be a historical negative trend in both the amount of phytoplankton and in the amount of plankton species that exist in these bodies of water (meaning decreasing biodiversity). In order to more clearly see this trend, a 5-year moving average was applied to the data and plotted below in **Figure 4.**

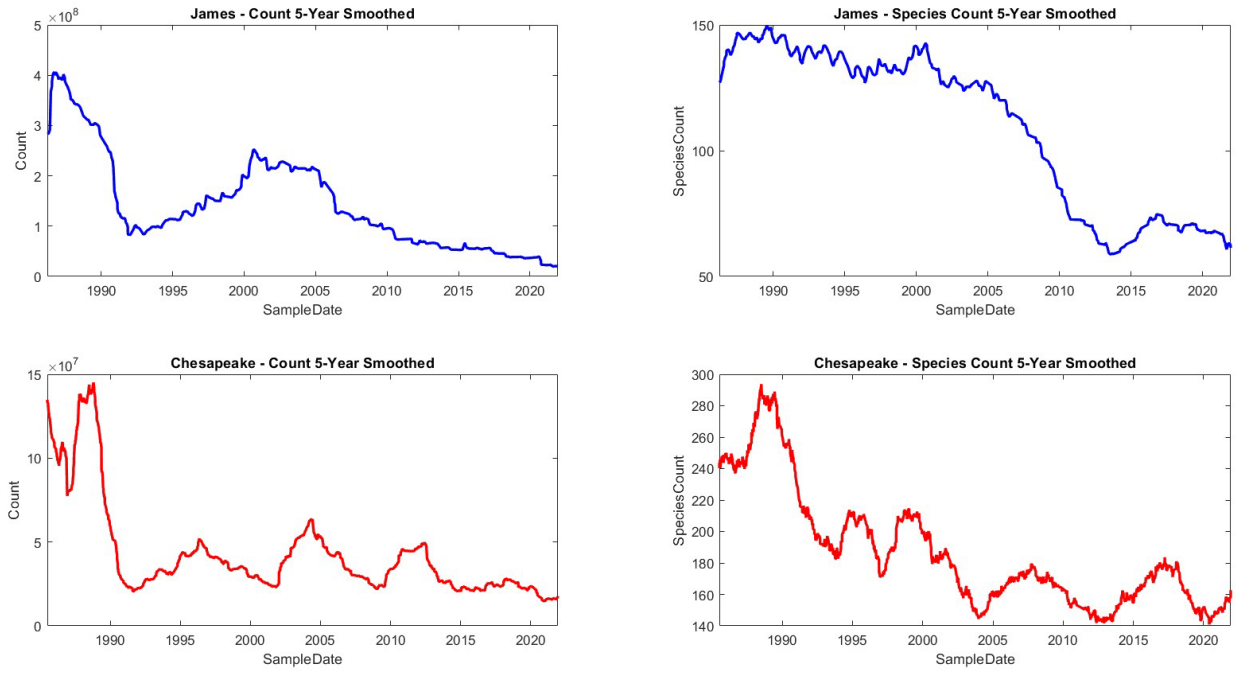

Figure 4: 5-Year Smoothed Plot of Phytoplankton Counts

We can now confirm that there have indeed been fairly significant decreases in both phytoplankton count and species count in both the Lower James and Lower Chesapeake. These observations allow us to hypothesize that changes in climate and water conditions over time have negatively impacted how phytoplankton have been able to subsist in their ecosystems. Specifically what might be driving these plankton counts down is explored in the following section.

2.3 Can water quality in the Lower James explain phytoplankton counts in both the Lower James and Lower Chesapeake?

In order to gauge what might be driving plankton counts down, I generated cross correlation function (CCF)[6] plots for each of the water quality features we previously explored with respect to plankton count and species count for both the Lower James and Lower Chesapeake. All data was transformed into stationary processes (when deemed necessary) in order to assure the assumptions of CCF's hold true. The non-stationary data and then data in which their first difference was applied[7], can be found in **Figure 8** and **Figure 9** in the appendix. The CCF plots can be found in **Figure 5** below.

[6] See https://online.stat.psu.edu/stat510/lesson/8/8.2 for more information on CCF's.
[7] https://otexts.com/fpp2/stationarity.html

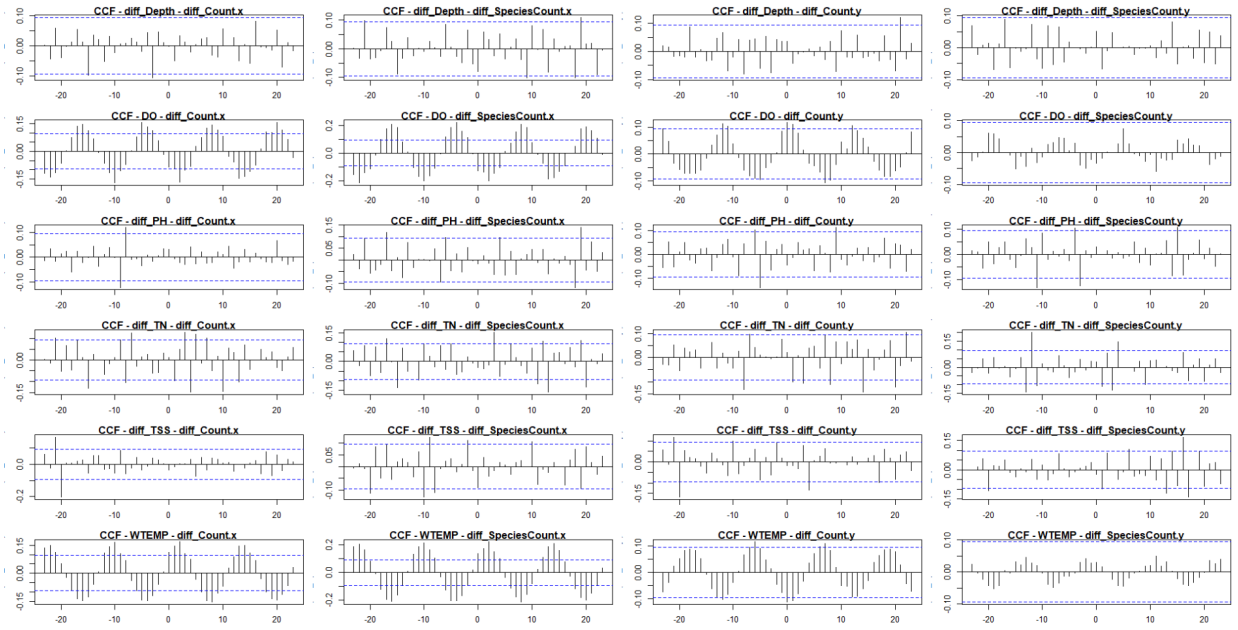

Figure 5: Cross Correlation Function Plots

Our plot may look intimidating at first (and a require a zoom in on your computer), but it be broken down pretty simplistically. First, looking at depth, the only really significant lags for all plots in this domain can be found for lag 18 and lag 19 (which signify 18 and 19 months, respectively, as this analysis was done with monthly averaged data) on Lower James Species Count and lag 21 for Chesapeake count, hinting that plankton counts are likely uncorrelated with water depths in the Lower James. With DO, most of our lags show a significant seasonal trend, highlighting that plankton counts are pretty much just historically correlated with DO at a seasonal level. The same exact explanation can be applied to water temperature as with DO.

When looking at pH, we begin to see something interesting. We see a variety of both significant positive and negative lags that highlight both positive and negative correlations. This observation makes sense as suitable pH for phytoplankton can almost be seen as a bell curve, where too low and too high of pH both can have detrimental effects on their vitality. A pH of 7 is the pH level that is “just right”. Significant negative and positive lags also tell us that phytoplankton counts are a product of past pH and can be likely explain future pH levels.

For total nitrogen (TN), we see a plethora of significant lags that tell us phytoplankton counts (both count and species count) are heavily correlated with TN. We see a presence of both positive and negative correlation at both positive and negative time lags. Our most significant lags, however, highlight that plankton counts are most likely to be positively correlated with nitrogen levels. **We have statistical evidence to say that if total nitrogen levels are “low” (low relative to historical nitrogen levels in the Lower James), then phytoplankton counts are likely going to become lower as well.** Again, there does seem to exist a balance of nitrogen levels for plankton, showing us that phytoplankton do indeed need SOME nitrogen to thrive, but too much (like through the introduction of pollutants to the water), can have negative effects.

Regarding total suspended solids (TSS), a measure of larger particles in the water (usually an indicator of how "clear" the water is), we see a variety of significant lags for all phytoplankton counts, with the most significant lags seeming to delineate a negative correlation with phytoplankton counts. This observation tells us that plankton species are likely to negatively effected by higher TSS in the water.

From this analysis, we can likely conclude that, of our variables, **total nitrogen levels explain variation in plankton counts the most**. Nitrogen is a vital nutrient to phytoplankton, but also includes nitrates/nitrites which are harmful, man-made pollutants, so a balance in nitrogen levels is key. The next most significant variables appear to be pH and TSS, with temperature and depth appearing to be more seasonally correlated than anything. **This analysis allows us to infer that phytoplankton counts have probably decreased due to the decrease of available nutrients (like Nitrogen), likely induced by the effects of climate change**.[8]

Chapter 3: The Future of The Lower James and Chesapeake Watersheds

To estimate how future phytoplankton species counts might look in the near future, I created and tested an autoregressive integrated moving average (ARIMA) model[9] that can be used for future forecasting. All features of interest that we explored earlier were used (see **Figure 10: Principal Component Analysis** in appendix for more detail). A plot of the data with the forecasted values, which were generated through fitting test data on our model, can be found below in **Figure 6**.

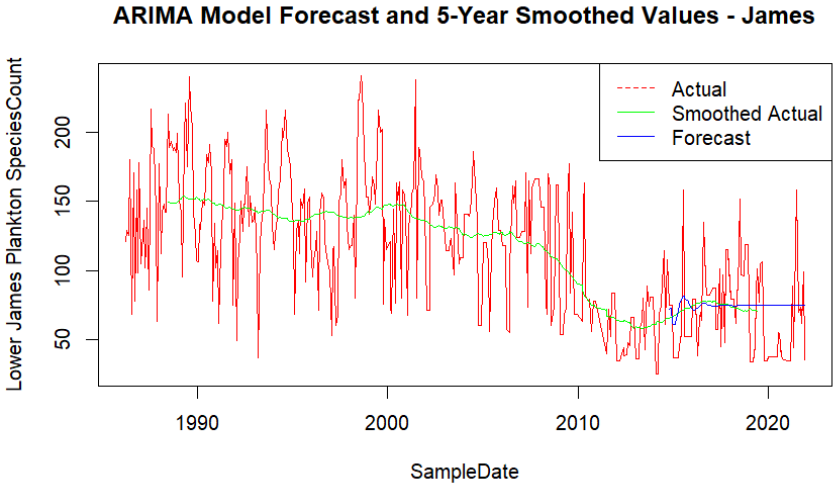

Figure 6: ARIMA Model for Species Count in the Lower James

Our model posts a test mean error of -4.09, meaning that, on average, our predictions for species count are only going to be around 4 units below the true value which is EXTREMELY powerful for a model of this nature. In depth accuracy results can be found in **Figure 11** the appendix.

[8] https://climate.mit.edu/explainers/phytoplankton#:~:text=Phytoplankton%20and%20climate%20change&text=Climate%20models%20suggest%20that%2C%20as,world%20will%20have%20fewer%20phytoplankton.
[9] https://people.duke.edu/~rnau/411arim.htm

A similar model was built to try and forecast phytoplankton species count in the Lower Chesapeake, based off of water readings from the Lower James. A plot of the data with the forecasted values, which were generated through fitting test data on our model, can be found below in **Figure 7.**

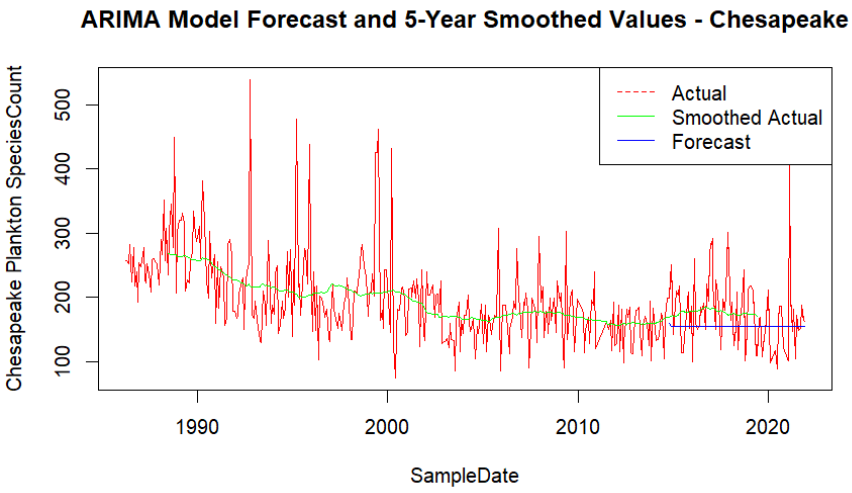

Figure 7: ARIMA Model for Species Count in the Lower Chesapeake

Our model posts a test mean error of 18.49, meaning that, on average, our predictions for species count are only going to be around 18 units higher than the true species count value which is fairly strong considering we are not even using variables recorded in the Lower Chesapeake. It is through this ARIMA model that we can further conclude that **water conditions in the Lower James DO have a significant impact on phytoplankton counts in the Lower Chesapeake**. Once again, in depth accuracy results can be found in **Figure 12** in the appendix.

Chapter 4: Conclusion

In conclusion, phytoplankton counts have significantly decreased over time in both the Lower James River and Lower Chesapeake Bay. Much of this decrease can be explained by variation in a variety of water condition/quality measures in the Lower James River like total nitrogen (TN), pH, dissolved oxygen (DO), water temperature, water depth, and total suspended solids (TSS). **The most significant of these variables appears to be TN which showed very significant (mostly negative) correlation with phytoplankton levels over time**. TN itself has seen a significant decrease over time in the Lower James, likely attributable to climate change.

We were also able to generate ARIMA models for both Lower James and Lower Chesapeake phytoplankton species counts which both posted strong accuracy results. One of the most major findings through this analysis was that **measures of water conditions/quality in the Lower James are able to explain a significant amount of variation in Lower Chesapeake phytoplankton species count**.

Appendix

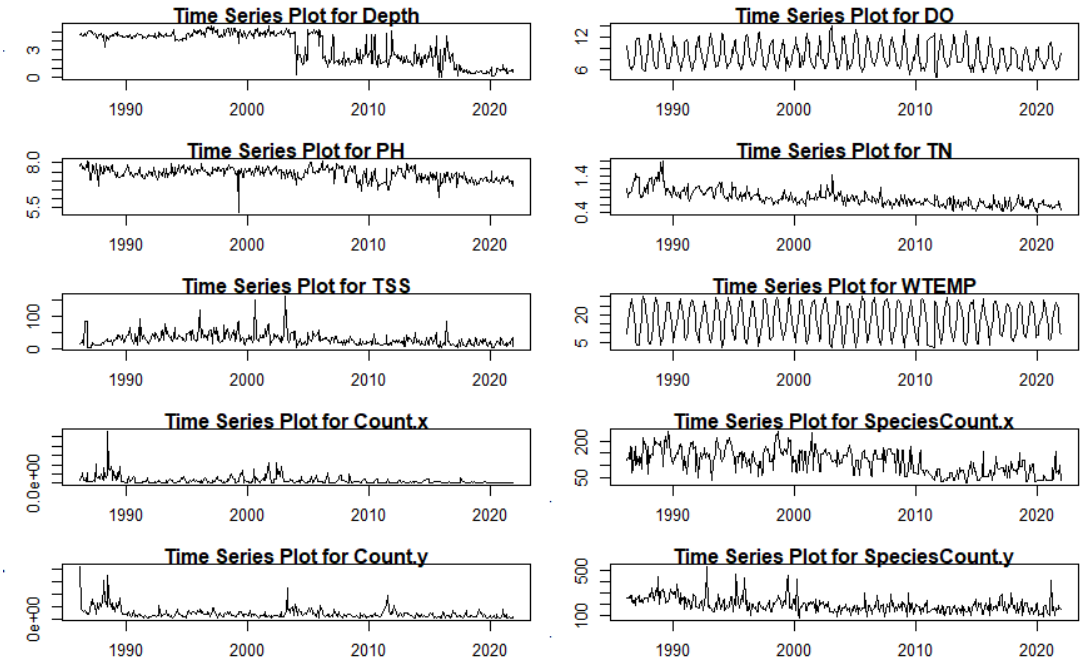

Figure 8: Data Pre-1st Difference

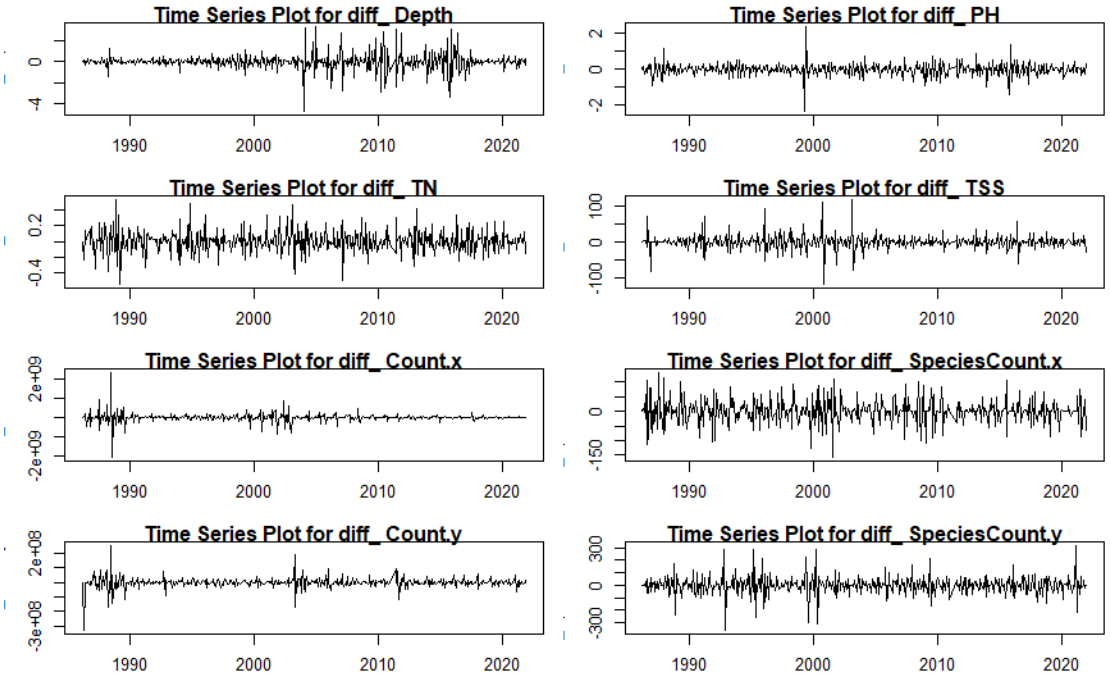

Figure 9: Data Post-1st Difference (where applicable)

```
Covariance Matrix:
[[ 1.00112486  0.17647218  0.74465753  0.45009703  0.63634317  0.0639264 ]
 [ 0.17647218  1.00112486  0.23175785  0.03497698  0.10145133 -0.89092077]
 [ 0.74465753  0.23175785  1.00112486  0.24319801  0.4343809   0.06387489]
 [ 0.45009703  0.03497698  0.24319801  1.00112486  0.45753992  0.03918223]
 [ 0.63634317  0.10145133  0.4343809   0.45753992  1.00112486  0.05791763]
 [ 0.0639264  -0.89092077  0.06387489  0.03918223  0.05791763  1.00112486]]

Eigenvalues:
[2.56740227 1.87741576 0.81505824 0.48304771 0.20103547 0.06278972]

Eigenvectors:
[[ 0.56691155  0.20254718  0.49415066  0.38673339  0.49183031 -0.0440601 ]
 [-0.0880697   0.67766984 -0.03585091 -0.11843921 -0.11208517 -0.71072575]
 [-0.20347504 -0.03772116 -0.56301154  0.76031307  0.20522466 -0.141421  ]
 [-0.06365934 -0.02160773 -0.33525432 -0.49004257  0.80081745 -0.040433  ]
 [-0.79081656  0.08391741  0.52925179  0.13429696  0.24768357  0.08987073]
 [-0.00309499 -0.7005757   0.2122379  -0.01249294  0.02769424 -0.68059984]]
```

Figure 10: Principal Component Analysis

This PCA output tells us a few things. First of all, we can see our eigenvalue vector's first two components are more than a magnitude of 1, telling us that our first two principal components (the first two columns of the Eigenvector matrix) explain most of the "important" variation in our data. A general rule of thumb is that any value over 0.5 in the eigenvector delineates a "significant" feature within the Principal Component. Also, it should be noted that each row in the eigenvector matrix represents the following features, respectively: Depth, DO, PH, TN, TSS, WTEMP. With this information, we can see that in PC1 (column 1), our most significant features are TSS, followed by Depth, meaning that these features explain the most "important" variation of our feature set. In PC2, we see that our next most important features include water temperature and DO.

Our third PC 's associated eigenvalue is just barely under 1 (our threshold for considering a PC), but we will go ahead and consider it anyway. It tells us that the third most important features in introducing variation in the data are pH and once again TSS. Our only variable not included in these three PC's is total nitrogen (TN). Because our exploratory data analysis showed such a significant trend for TN over time, I decided to go ahead and include it. This PCA ended up turning out to be not super informative, but still a necessary exercise for creating a predictive model that is as informative as it could be, hence the reason for being in the appendix.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-2.62027	56.32230	38.34246	-8.082606	20.98280	0.8077014	0.0003348414
Test set	18.48568	57.95582	42.90345	2.058988	24.35865	0.9037809	NA

Figure 11: Accuracy Results for James Species Count ARIMA Model

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.5282821	34.22030	25.77700	-10.26585	27.19618	0.9249283	-0.01228187
Test set	-4.0850239	31.45069	26.57962	-28.62495	48.59934	0.9537279	NA

Figure 12: Accuracy Results for Chesapeake Species Count ARIMA Model