**Predicting Future Ocean pH**

Danielle Lamb

Bellevue University

DSC 630-T302

**Executive Summary**

The health of the oceans is very important to the health of the planet overall. A general rule of thumb for ocean health is pH and the health of coral reefs. The measure for whether something is acidic or alkaline is called pH. This is a measure with a range of 0 to 14 with 0 being most acidic and 14 being most alkaline. Stomach acid is very acidic with a pH of between 1.5 and 3.5, while lye, a common ingredient in household drain and oven cleaners is extremely alkaline and has a pH of 14. The ocean has an average pH of 8.1 and does not handle extreme changes in pH very well.

The pH of the oceans reflects the amount of carbon dioxide in the water and air. With companies producing more carbon dioxide with creating goods, shipping, and other processes, the ocean absorbs more of this carbon dioxide which could cause pH to decrease.

A Long Short Term Memory Network was created with existing pH data to create a program that can predict the value of pH. By using this model to predict future pH, one can detect when the ocean pH may become too acidic for most ocean organisms to live. Then, a timeline for new processes that include lower carbon emissions for companies and citizens can be created to prevent ocean pH from becoming too acidic to the point of decreased oxygen production. If a process such as this is not implemented, and oxygen levels are severely decreased because of lowering numbers of oxygen producing microorganisms, a global disaster could happen.

Currently this model was able to predict future pH in data that was included in this study. More processes and further understanding would be able to produce a forecast into the future much like a weather forecast.

**Abstract**

A majority of the Earth’s oxygen is produced by photosynthesizing microorganisms called phytoplankton that reside in the surface ocean, between 600 and 900 feet, the euphotic zone. To protect themselves, the plankton build calcium carbonate shells. Calcium carbonate is very susceptible to pH changes, low pH can cause calcium carbonate to dissolve. Data collected from a mooring in the Pacific Ocean will be analyzed with a Vector Autoregression Model as well as a Long Short Term Memory Network. These models will be able to learn from historical data ocean conditions that indicate a lowering pH. The ideal model was the Long Short Term Memory Network with a testing R2 of 0.77. This will help predict future pH in order to understand the efforts large companies and ocean related businesses need to make to help reduce their impacts of lowering pH. Understanding the trend of pH as years pass can help in allocating resources for reducing carbon emissions and improving ocean health.

**Background**

The Earth commonly has the nickname “The Blue Planet” because there is a higher percentage of water and oceans than land. All of the world’s oceans are interconnected in different ways, and generally have similar compositions to each other. The way acidity and alkalinity is measured in water is called pH. The scale of pH is between 0 and 14. “Typically, the surface waters of today’s ocean have a pH of around 8.1,”1. A neutral pH, meaning not acidic or alkaline, is 7.0, so generally the ocean is alkaline to a degree. This value has only varied by around 0.1 over the past hundred years. Recently, the pH of the oceans has been changing at approximately 10 times the rate than the last several million years. “Ocean acidification is an ongoing large-scale environmental problem, whereby the absorption of anthropogenic CO2 by the ocean lower its pH, impacting ocean ecosystems worldwide,”2.

Carbon is the basis of all life on earth, including marine organisms. Numerous of these organisms have a calcium carbonate shell that surrounds it. A few large examples are corals, which have a symbiotic relationship with a tiny algae called zooxanthellae, zooplankton, and phytoplankton. These phytoplankton are some of the most important organisms on Earth. They can account for approximately 70% of total oxygen production. So, a majority of the oxygen in the atmosphere comes from these photosynthesizing plankton in the ocean. With decreasing pH, and increasing amounts of carbon dioxide in ocean water, the ability to grow calcium carbonate for shells, bodies, and corals are impeded. In addition, current calcium carbonate is more easily dissolved with additional carbon dioxide and lower pH. With this, oxygen production in the future could decrease since the phytoplankton cannot efficiently create shells or their shells dissolve away.

Current research has used forecasting models similar to weather forecasting models. These models are fairly new and have not been widely used. “They found that the climate model forecasts did an excellent job at making predictions of ocean acidity in the real world,”3. In addition, smaller, more focused areas have made their own models that assist fisheries. With the development of this larger scale model that can factor in global climates like El Niño years, the smaller models could benefit significantly in improving their forecasts.

**Methods**

Data from mooring Chuuk K1 within the National Oceanic and Atmospheric Administration (NOAA) was collected starting in 2011 and ending in 2018. This mooring is located off the coast of Micronesia in the Western Pacific ocean. Null values and values impossibly extreme (-999 or 999) were removed. Redundant features including mooring name, latitude, and longitude were all removed since each instance had the same value. To determine which features are related to the target, a correlation matrix was created and a heat map to visualize the results. Those features not correlated were removed. The date and time was converted into index. A pairplot was created to visualize relationships. Boxplots of each feature were also created to find outliers. Histograms were created to evaluate normality of features. Lineplots were made to analyze trends. Seasonality trends must be addressed before populating prediction models. The target feature, pH, was grouped by time periods of day, week, month, quarter, and year by average to further visualize trend.

A vector autoregression model was the first attempt at the predictive model. The dataset was split into training and testing by a 80/20 split. Next, a Johanson’s Cointegration test was performed to evaluate if three or more time series are correlated long term. After, an Augmented Dickey-Fuller test was performed to test for stationarity of features. The first difference was taken on all features to produce stationarity. To find optimal lags, the VAR was taken over 10 numbers, with highlights on lowest values for best lag. The model was fitted with a lag of 3, and residuals were assessed. A forecast was created and values were inverted from 1st difference and graphed. An additional graph of all features against actual values was created as well as a specific visualization of the target feature, pH.

Next, a Long Short Term Memory Network (LSTM) was created as an additional model. This model used numerous keras and tensorflow libraries. This model is a single feature model, so pH was taken from the original dataset and made into an array. The shape of the data was then corrected and split into training and testing sets with the same 80/20 split. Originally, a lookback of 1 was used. An LSTM was fit using 100 epochs at batch size 1. The transformations were inverted and RMSE for both training and testing sets were calculated. A lineplot with the data, training predictions, and testing predictions were plotted for visual results. Lastly, an R2 was calculated for both training and testing sets.

**Results**

Zero null values and zero duplicates were found in the data. There were numerous features with no significant relation to pH, most include the measurements of carbon dioxide in the air. Boxplots showed that only atmospheric pressure had two outliers. The histograms resulted in many of the features having non-normal distributions. Since there is a degree of seasonality, these distributions were left alone. Lineplots indicated seasonal increases and decreases in all features. Maximum peaks occurred during July and August of each year. Minimum peaks occurred around January and February of each year. The trend of pH was inverted from all other features. The detailed plots of pH showed a general negative trend from July through December, and a general positive trend from January to June. The yearly trend was negative at approximately 45 degrees.

The Johanson’s Cointegration test showed a lasting correlations of concentration of CO2 in seawater (wet) and atmospheric pressure. The first Augmented Dickey-Fuller test resulted in numerous features being non-stationary. The 1st difference was taken and the second run of the Augmented Dickey-Fuller test resulted in all stationary features. The VAR resulted in an ideal lag of 3. When looking at the maximum and minimums of the forecasted data, the range is much more narrow than the actual data.

The LSTM had a 0.01 RMSE for both testing and training scores. The visualization resulted in a much closer prediction than the VAR. The LSTM had a training R2 of 0.92 and a testing R2 of 0.77.

**Discussion**

Trends of all the features tend to increase from March through June, peak around July and then decrease through December. This is due to the absorbability of the water at different temperatures. In the cooler months, the water has a higher ability to absorb more things including carbon dioxide, which is why there is a trend into the warmer months. Once that temperature reaches a peak around July, the ability to absorb decreases, which is why there is a decreasing trend into the winter months. With a higher concentration of carbon dioxide (such as xCO2 in the data) the pH is driven lower, giving it an inverted correlation. In addition, the low spike of pH in 2014 when compared to 2013, is lower. This trend is supported by the yearly average figure as well. Additional data would be necessary to confirm a yearly trend of lowering minimum pH, indicating more acidic water.

The Johanson’s Cointegration test indicated the concentration of carbon dioxide, licor atmospheric pressure reading, and sea surface temperature both had long term correlations with at least three other features. Some of the other features are partially measured off of the concentration value, which could be a cause for this result. Further tests would have to be performed to narrow down which features have long term correlations with the atm pressure, one from the correlation matrix is concentration of carbon dioxide in the air.

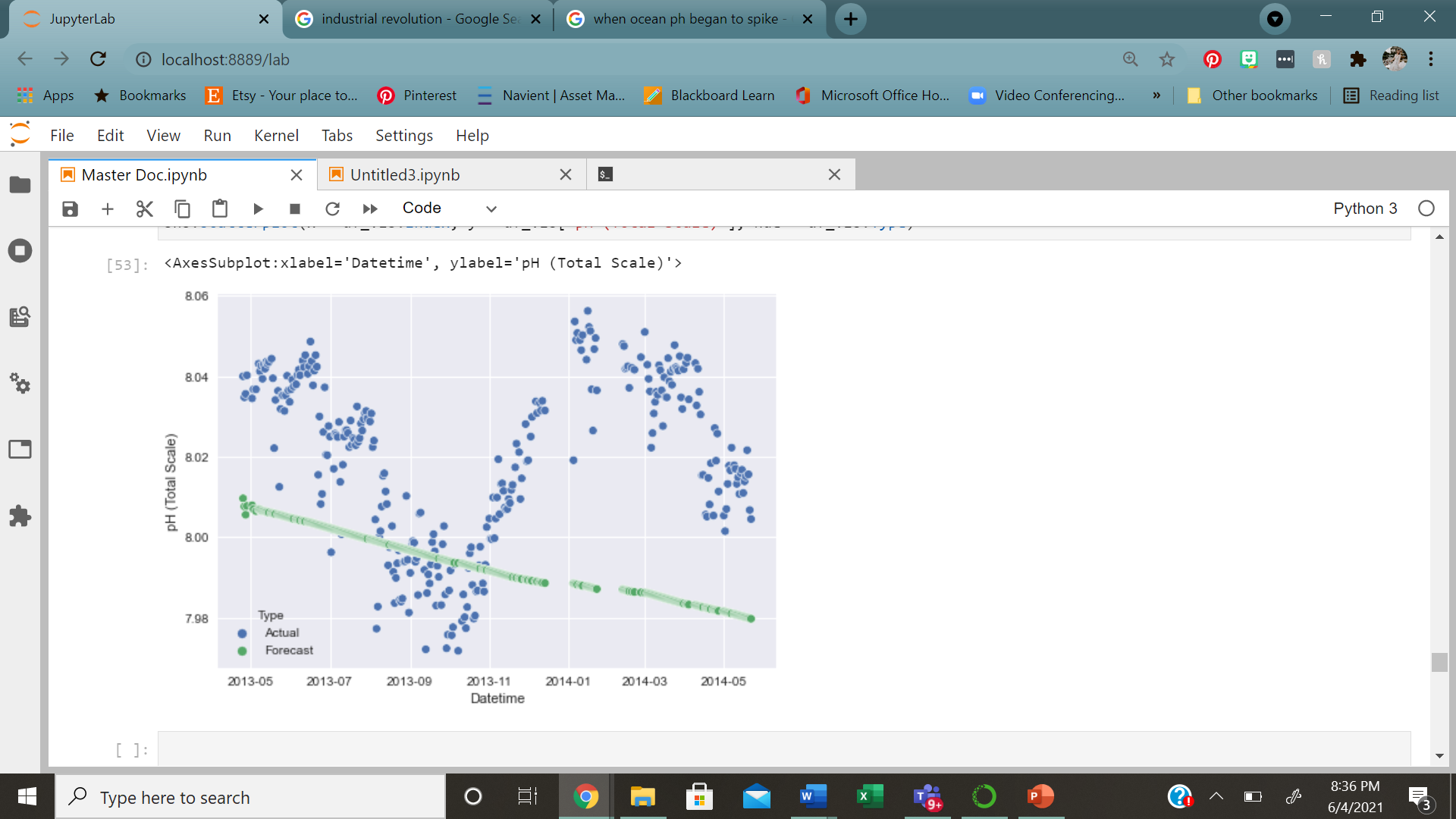


Figure 1: Actual data vs. Forecasted data using vector autoregression model.

Overall the vector autoregression model was not a good fit for this data. The model was able to predict an overall trend, but not specific pH values. With the addition of the other features, this model may have had too many inputs to focus on a correct prediction for one value. All other features had similar, inconclusive results in prediction.

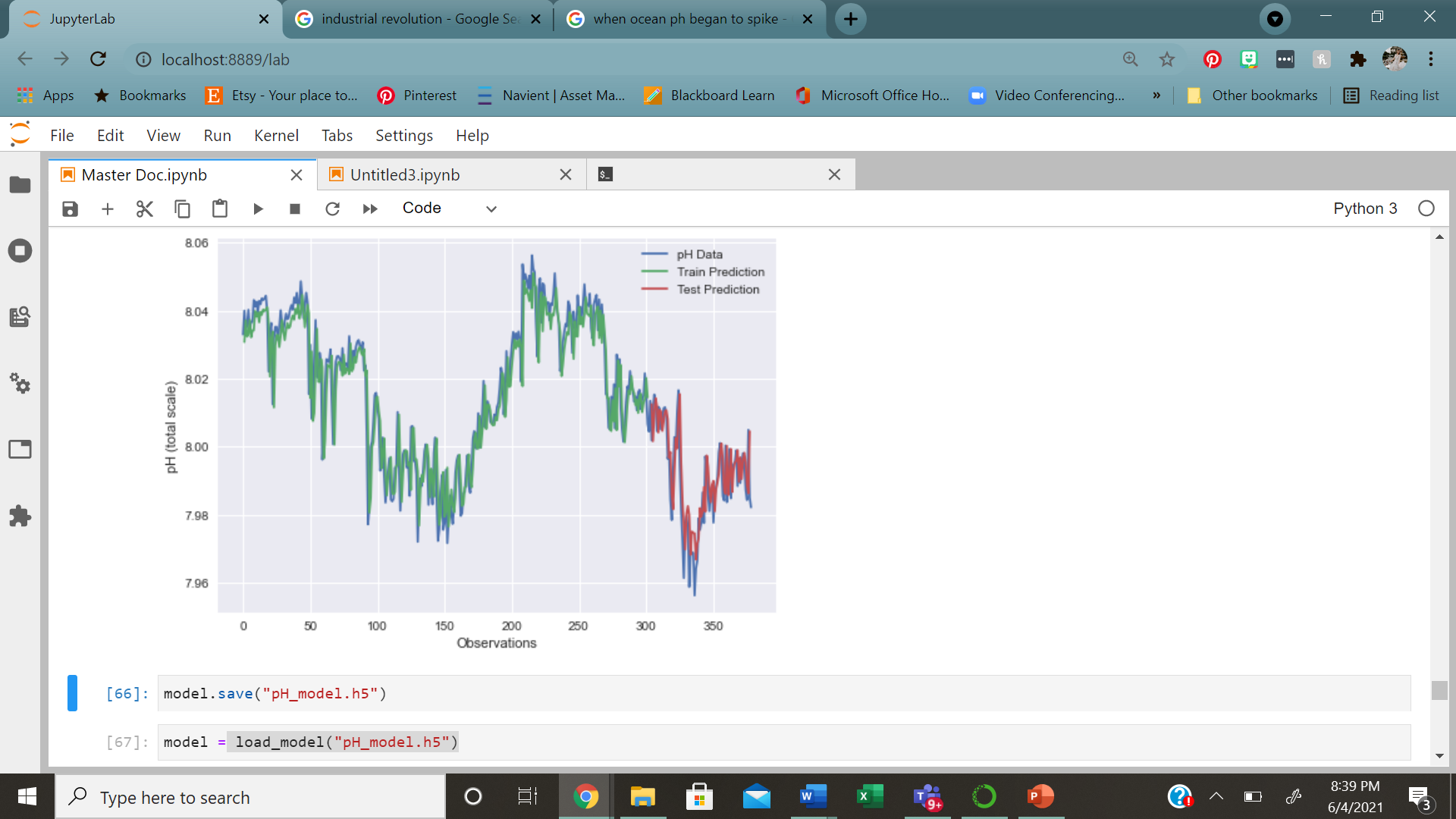


Figure 2: Results from LSTM model. Green indicates training predictions and red indicates testing predictions.

The second model was the Long Short Term Memory Network (LSTM). This model used only pH as an input, along with the time series. This model was highly accurate when compared to the VAR model used previously. When testing the model, it proved to be accurate in predicting the pH 77% of the time. More data introduced could help this model become more accurate to upper and lower extremes.

**Acknowledgements**

The author would like to thank the Marine Science department at Coastal Carolina University for inspiring a passion for ocean acidification and ocean health. Additionally, the author thanks Professor Werner and the Bellevue University Data Science department for providing a program to teach students how to maximize their understanding of statistics and data analysis. Lastly, the author would like to thank Courtney Klatt and Billy Andrews for their support through peer reviews and correspondence on the project.

**References**

Barker, S. & Ridgwell, A. (2012) Ocean Acidification. *Nature Education Knowledge* 3(10):21

Brady, R.X., Lovenduski, N.S., Yeager, S.G. *et al.* (2020) Skillful multiyear predictions of ocean acidification in the California Current System. *Nat Commun* 11**,**2166.

University of Colorado at Boulder. (2020) "Ocean acidification prediction now possible years in advance." ScienceDaily.