Tracking Climate Change and Sea Surface Temperatures with Time Series and Machine Learning
By Elena Eguiarte

Increasing ocean temperatures have a severe effect on marine species and ecosystems. The ocean absorbs most heat from greenhouse gas emissions, leading to rising ocean temperatures. Data from the US National Oceanic and Atmospheric Administration (NOAA) shows that the average global sea surface temperature has increased by 0.13 °C per decade over the past 100 years. Time series and machine learning modeling are useful techniques to predict and understand what contributes to water temperature increase. Understanding the seasonal trends and features affecting sea surface temperature helps indicate what contributes to the increase in water temperature. Using Machine Learning and time series modeling will allow us to mitigate rising temperatures promptly.

For this project, a total of three datasets were acquired. The first dataset is the El Niño dataset, which contains oceanographic temperature recordings taken from a series of buoys positioned throughout the Pacific Ocean from 1980 to 1998; this dataset was obtained from Kaggle, but the recordings were collected by NOAA. The other two datasets are from the California Cooperative Ocean Fisheries Investigation (CalCOFI). The second dataset contains oceanographic data measured from seawater samples collected at CalCOFI stations; these recordings include dissolved oxygen, salinity, and nutrients found in the ocean. This dataset is known as the "Bottle" database. The third dataset, known as the "Cast" database, contains metadata that includes date, time, latitude, and longitude. The CalCOFI recording is based on the Northeastern Pacific Ocean area. The CalCOFI datasets recordings are from 1949 to 2020.

The first dataset that was cleaned and preprocessed was the El Niño dataset. This dataset originally has 178,080 rows and 12 columns and was used for the time series model, and the target variable is "Sea Surface Temp." Around 10% of missing values were for the "Sea Surface Temp" column. For time series modeling, all dates need to be unique. In this case, there were no duplicate dates. However, there were some duplicates in the Latitude and Longitude columns. The best way to solve this was to obtain the median of these coordinates; from there, a backward filling

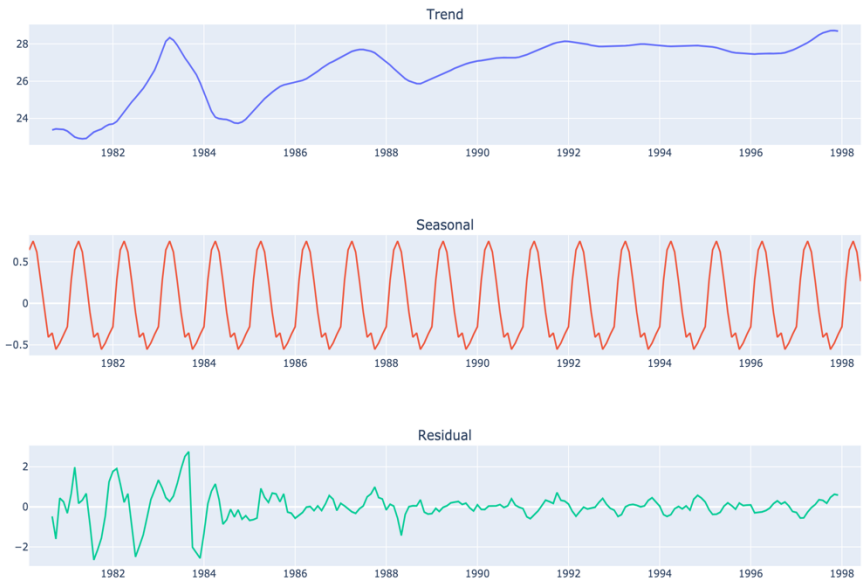

Figure 1.1 Trend-Seasonal Decomposition

was done for all the sea surface temperature null values. Backward filling fills missing values with the next non-null value. Backward filling was done because most of the missing values for the "Sea Surface Temperature" are from 1986 to 1990; therefore, it is the best option for the next non-null value to be filled rather than the previous non-null value. The dataset was then decomposed, as shown in Figure 1.1. After the decomposition, the seasonal differences were obtained to make the time series stationary. Then the Seasonal AutoRegressive Integrated Moving Average (SARIMA) was used; SARIMA specifically accounts for seasonality by incorporating seasonal differencing and seasonal auto-regressive and moving average components. Figure 1.2 is SARIMA demonstrates the train data, which spans from 1988 to 1996, and the testing data which spans from 1997 to 1998. The predictions demonstrate a decrease in Sea Surface Temperature but sharply increases.

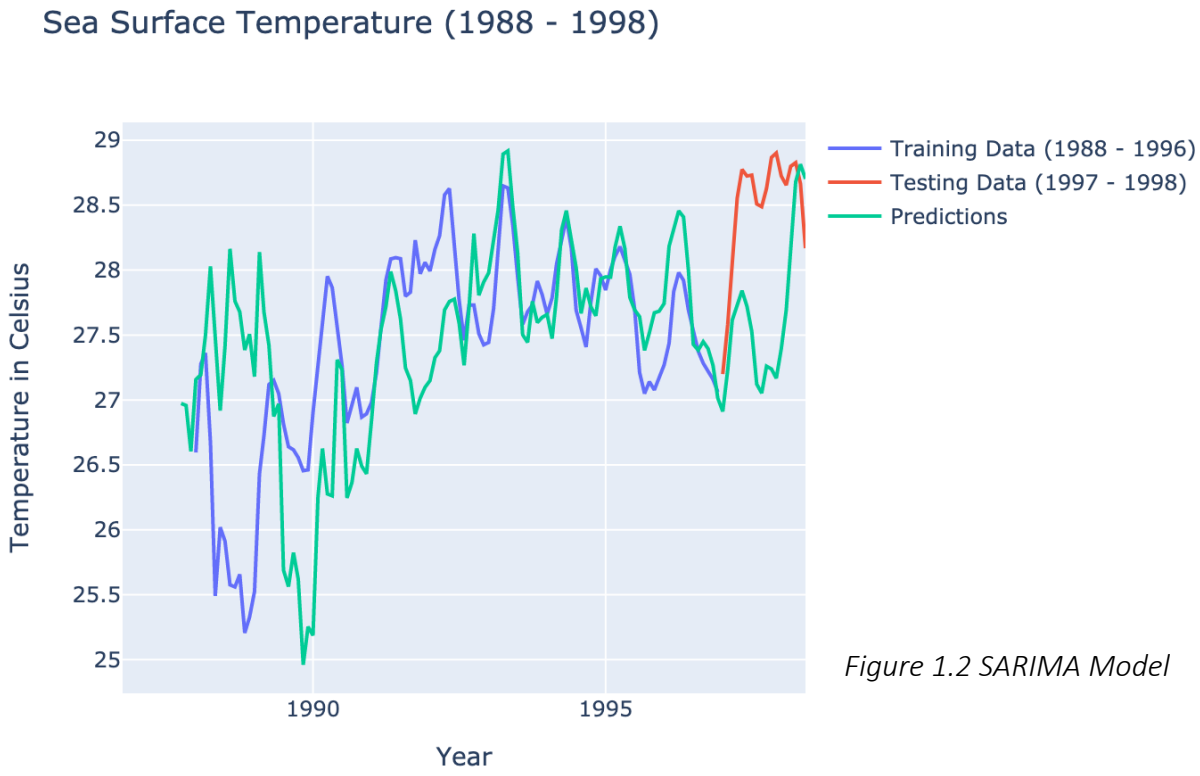

Figure 1.2 SARIMA Model

For the next dataset, both bottle and cast datasets were explored. The bottle dataset data frame originally contained 889,500 rows and 62 columns. First, the correlation between water temperature (°C) and the other features was analyzed. Then the columns that had a high correlation with water temperature were selected. Some of the selected columns were dropped because 99% of their values were null. Afterward, the cast database was analyzed. The bottle and cast datasets were merged using the Cast Count (Cst_Cnt). The final merged data frame contains the following columns:

- Potential Density (Sigma Theta), Kg/M³ (STheta)
- Micromoles of Phosphate per liter of seawater (PO4uM)
- Micromoles Nitrate per liter of seawater (NO3uM)

- Micromoles Silicate per liter of seawater (SiO3uM)
- Water temperature in degrees Celsius (T_degC)
- Cast Count (Cst_Cnt)
- Date, Latitude, and Longitude (Lat_Deg and Lon_Deg)

After both data frames were successfully merged, a linear regression model was initiated. First, some diagnostics were done in order to ensure the results are reliable and valid. The first assessments were linearity and multicollinearity - this was done before running the model. Linearity checks the relationship between X (predictors) and y (target variable). Reported Salinity (R_Sal) is the only feature that showed positive linearity with temperature. As for multicollinearity, Variance Inflation (VIF) was used to quantify the severity of multicollinearity; multicollinearity occurs when predictor variables in a regression model are highly correlated. After running the Linear Regression Model, an R^2 value of 0.89 was obtained. This indicates that 89% of the differences in water temperature using oxygen concentration, salinity, and depth. In other words, this model can predict water temperature based on these three factors. However, there is 11% of the changes in water temperatures that this model cannot explain. After running the linear regression model, normality and homoscedasticity were assessed. The distribution of residuals was normally distributed. As depicted in Figure 1.3, the homoscedasticity briefly demonstrated some heteroscedasticity, but eventually, the residuals showed an equal variance. The third model was a decision tree regression model; the R^2 value for this model was 0.94, meaning that the decision tree regression model can explain 94% of the differences in water temperatures using the other features. This model can predict water temperature based on the features eight selected features. Similar to the linear regression model, a residual plot was done and demonstrated a similar plot to the one in Figure 1.3.

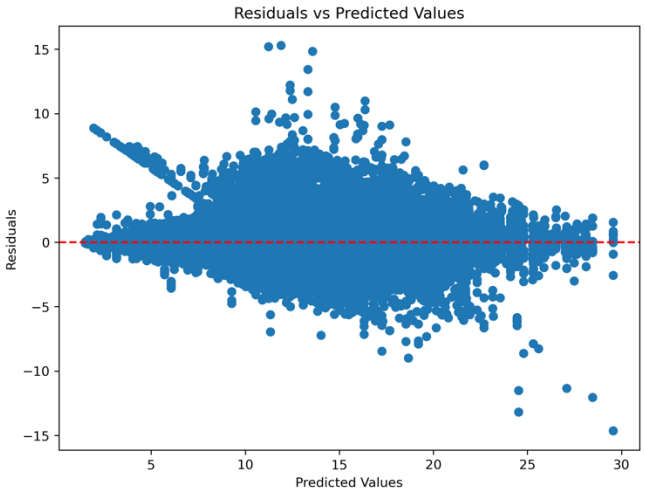

Figure 1.3 Homoscedasticity

This study aimed to predict water temperature based on eight features. Both models were able to predict water temperature reasonably well using the provided features (in the case of linear regression, only three features were able to predict sea temperature). While both models demonstrate strong predictive capabilities, additional evaluation methods should be employed to ensure the models' robustness and generalizability to unseen data. Additionally, particular attention should be paid to the potential overfitting issue in the Decision Tree Regression Model. Whereas the time series, a successful SARIMA was employed, showing a Mean Absolute error (MAPE) of 3.49% on the train set and a MAPE value of 6.82% in the test set.

References

Baloğlu, B. (2020, June 11). *Tackling Climate Crisis with machine learning*. Medium. Retrieved April 9, 2023, from https://towardsdatascience.com/tackling-climate-crisis-with-machine-learning-d9426fe1f5a9

Bottle Database. CalCOFI. (n.d.). Retrieved April 9, 2023, from https://calcofi.org/data/oceanographic-data/bottle-database/

Learning, U. C. I. M. (2016, November 6). *El Nino Dataset*. Kaggle. Retrieved April 9, 2023, from https://www.kaggle.com/datasets/uciml/el-nino-dataset

Ocean warming. IUCN. (2022, July 20). Retrieved April 9, 2023, from https://www.iucn.org/resources/issues-brief/ocean-warming