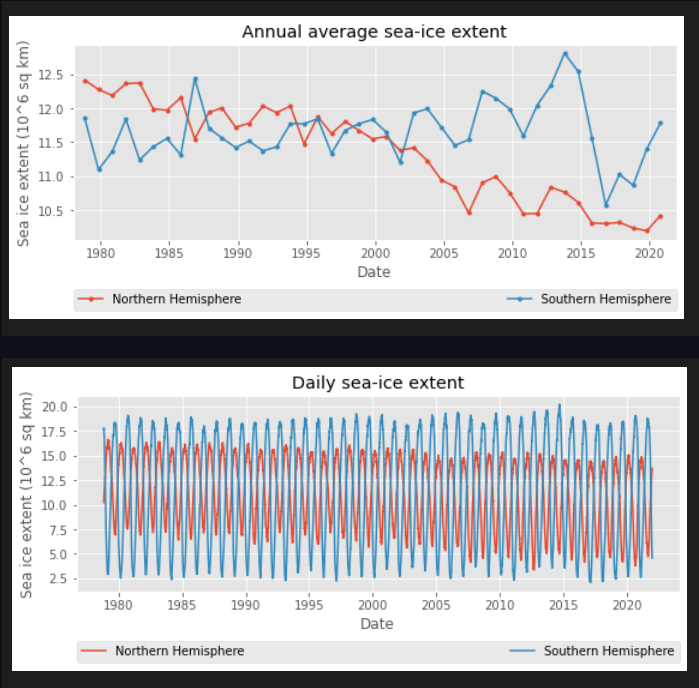
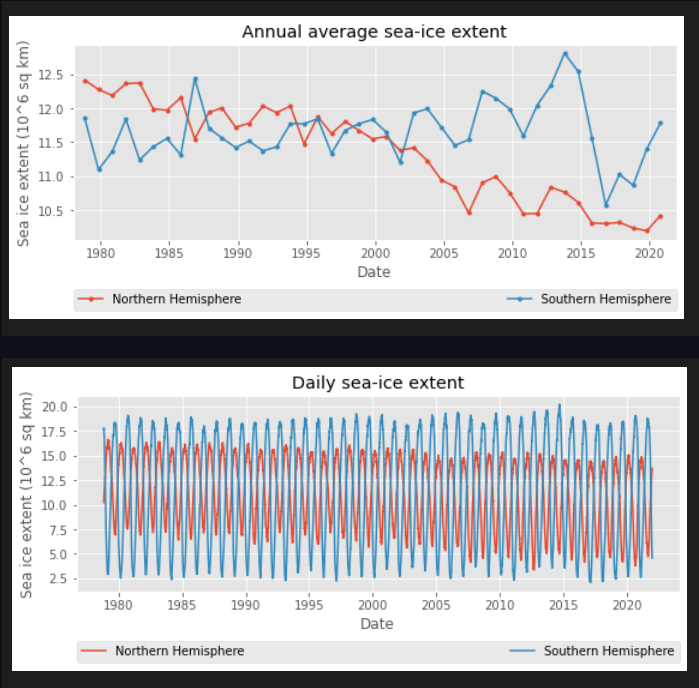
**Data Exploration / Gathering**

Based on research, we discovered the following trends in the sea ice extent size between 1980 and 2021:



Machine Learning - Model Training



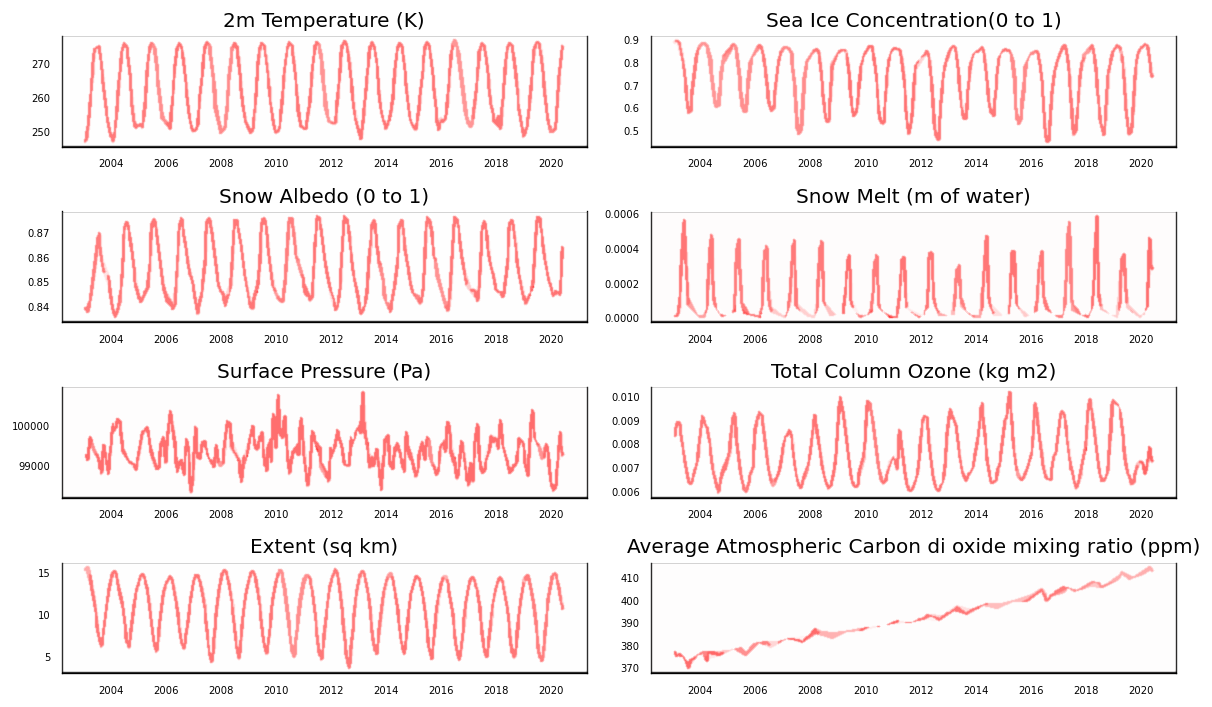
In order to identify which factors, contribute to this trend, we reviewed data from the following resources and identified the most useful features to be used in our analysis:

* National Snow and Ice Data Center (NSIDC) (https://nsidc.org/data/google\_earth)
* Climate Data Store (https://cds.climate.copernicus.eu/user/119111)
* Visualize Arctic and Antarctic Sea Ice (https://livingatlas.arcgis.com/sea-ice/)

We extracted the following information:

|  |  |  |  |
| --- | --- | --- | --- |
| Source | Table Name | Unit Measurement | Definition |
| NSIDC | sea ice extent (Artic) | sq km | Sea ice extent is the area of ice that covers the Arctic Ocean at a given time |
| Copernicus | 2m temperature | K | This parameter is the temperature of air at 2m above the surface. |
| Copernicus | Sea-ice cover | (0-1) | Area covered by sea ice. Long-term monitoring of sea ice is important for understanding climate change. Sea ice also affects shipping routes through the polar regions. |
| Copernicus | Snow albedo | (0-1) | Measure of the reflectivity of the snow-covered area. Solar (shortwave) radiation reflected by snow across the solar spectrum. |
| Copernicus | Snowmelt | m of water | This parameter is the accumulated amount of water that has melted from snow in the snow-covered area |
| Copernicus | Surface pressure | Pa | Pressure of the atmosphere at the surface |
| Copernicus | Total column ozone | kg m-2 | Total column ozone is the total amount of ozone oon the earth's surface to the top of the atmosphere. It is measured Dobson units (DU). The ozone hole is defined in terms of reduced total column ozone—less than 220 DU. |
| Copernicus | Average atmospheric carbon dixide (XCO2) | ppm | Average number of XCo2 molecules in the surface area of air. |
| Copernicus | Average atmospheric carbon dioxide (CO2) | ppm | Average number of Co2 molecules in the surface area of air. |

The trends of these features were assessed:



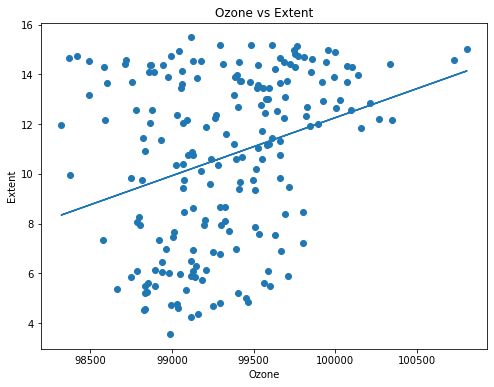
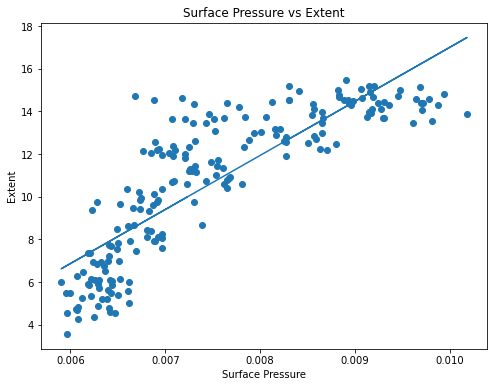
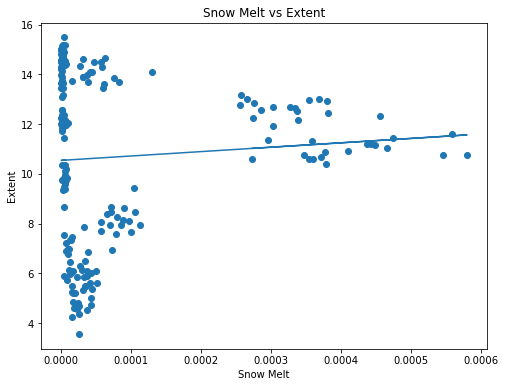
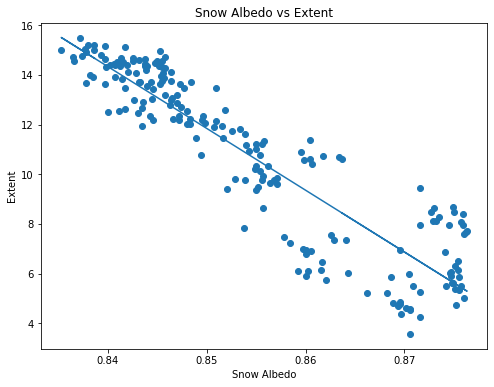
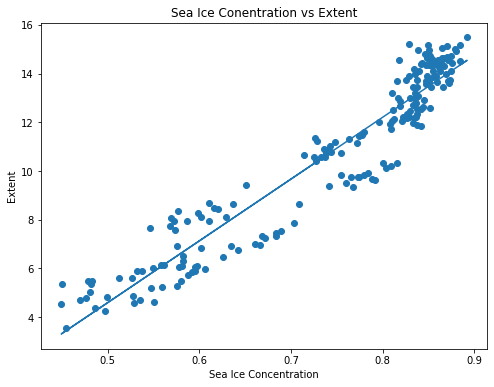
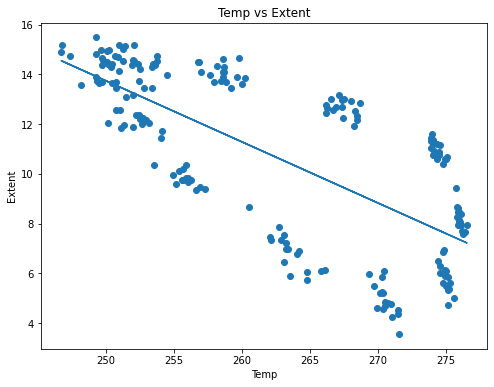
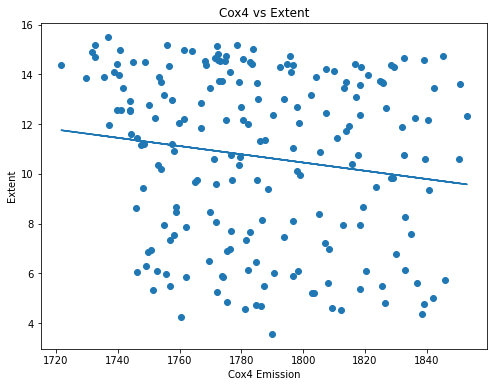
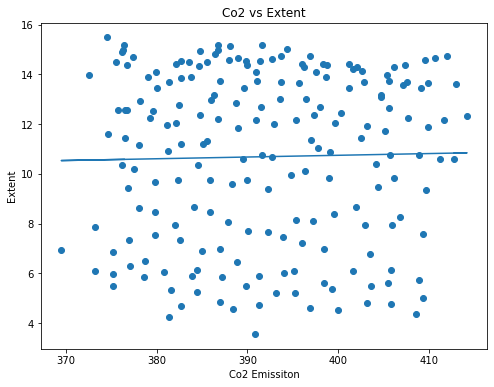
(LINK TO THE DATASETS FROM GITHUB)

<https://github.com/ALIYA2Group/Mod20_Segment_3/blob/main/ETL%20Data/N_seaice_extent_daily_v3.0.csv>

<https://github.com/ALIYA2Group/Mod20_Segment_3/blob/main/ETL%20Data/North_Data_Finalized_1.csv>

**Regression Analysis**

The selected features were visualized as time series and against the target (extent) to understand the correlation.



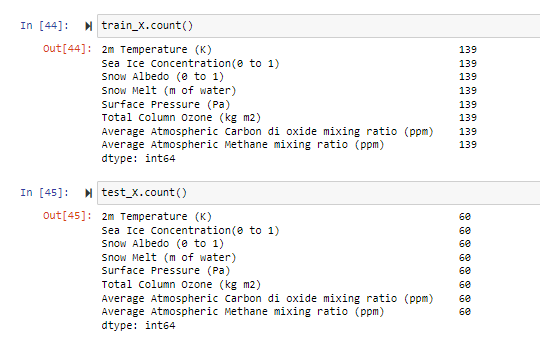
**Machine Learning – Model Training**

Code available at:

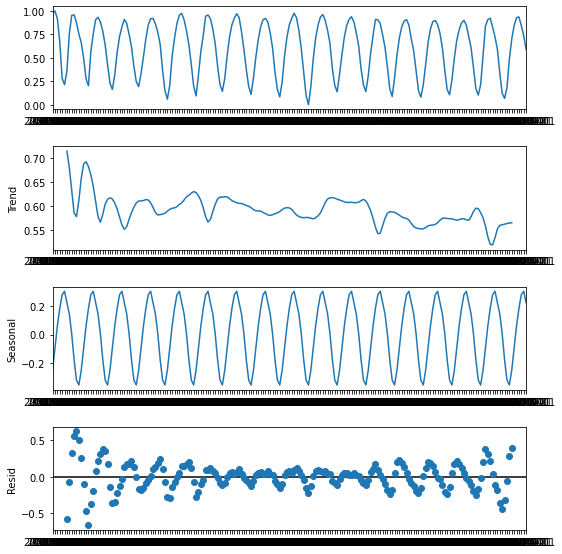
https://github.com/ALIYA2Group/Mod20\_Segment\_3/blob/main/Machine%20Learning/ML\_Model\_North\_SARIMAX.ipynb

Given the nature of the data and the question we are trying to answer, we used a Timer-Series prediction model **SARIMAX** (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors). It helps to predict future values using auto-regression and moving average along with adding in the seasonality factor.

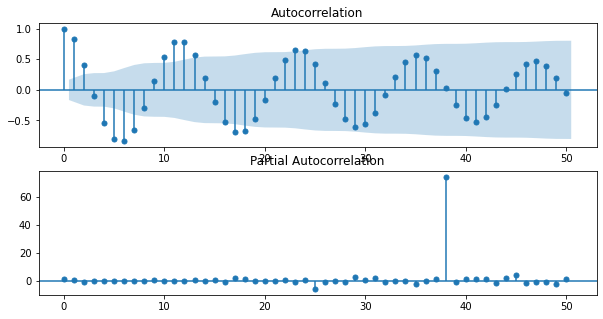
* Data pre-processing included importing the dataset using SQLAlchemy from AWS, dropping unwanted columns and setting the date as index. Data was also scaled using MinMax Scaler from the Scikit library.
* Data was split into training and testing sets using a 70-30 ratio and using the scikit library.



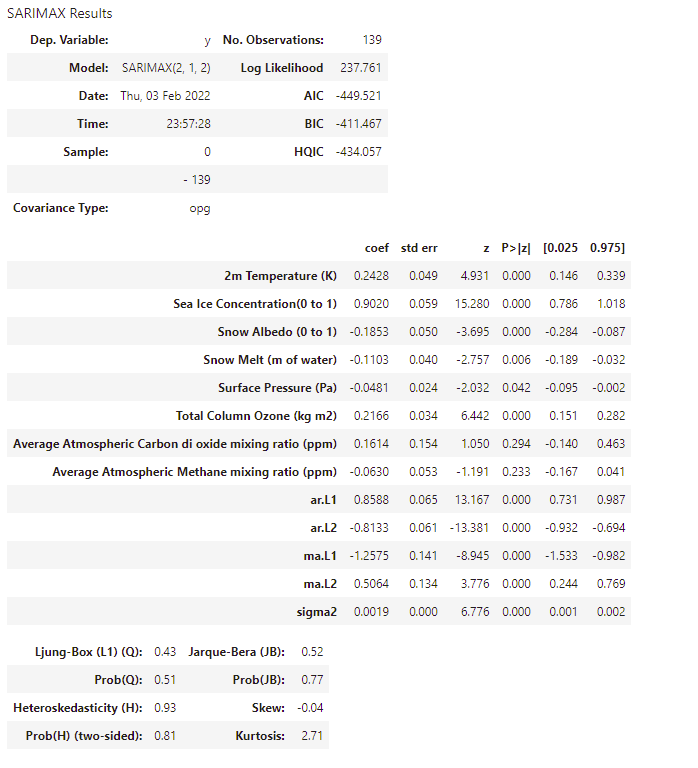
* Decomposed time-series into several components – trend, seasonality, random noise



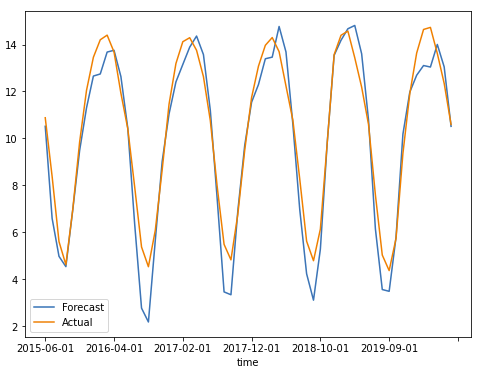
* Checked for Data Stationarity using Augmented Dickey-Fuller(ADF) test. If we make the data stationary, then the model can make predictions based on the fact that mean and variance will remain the same in the future. A stationarized series is easier to predict. For data points that were not stationary, data was differenced to make it stationary.
* An ACF and PACF bar chart was plotted. ACF is a plot of the coefficients of correlation between a time series and its lag and helps determine the value of p or the AR term while PACF is a plot of the partial correlation coefficients between the series and lags of itself and helps determine the value of q or the MA term. Both p and q are required input parameters for the SARIMAX Model.



* Using the pyramid and statsmodel libraries, Ran the SARIMAX model to forecast the extent based on the order obtained using ARIMA model and using the training set as the exogenous variables



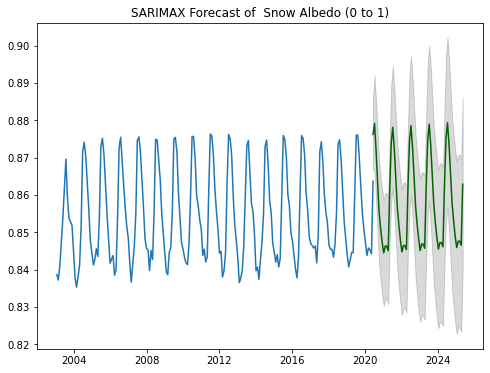
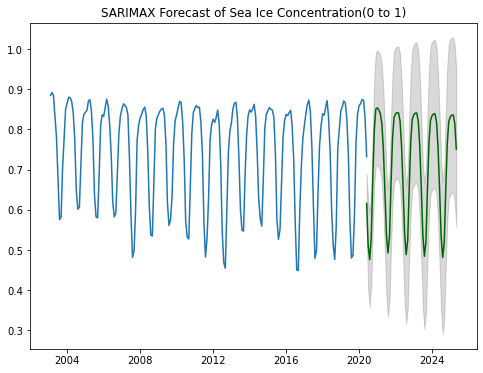
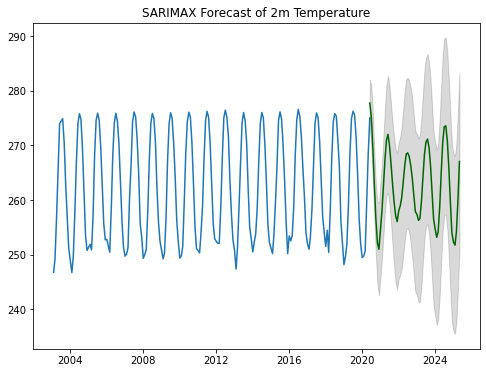
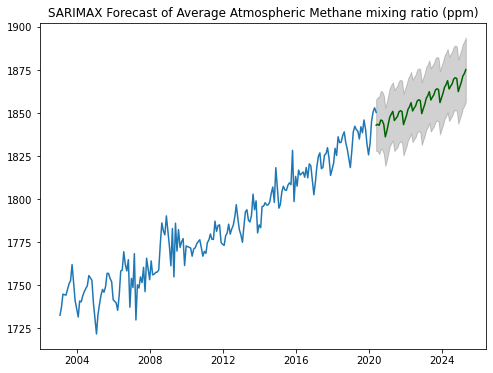
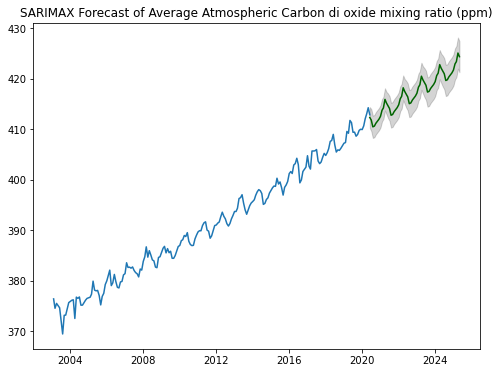
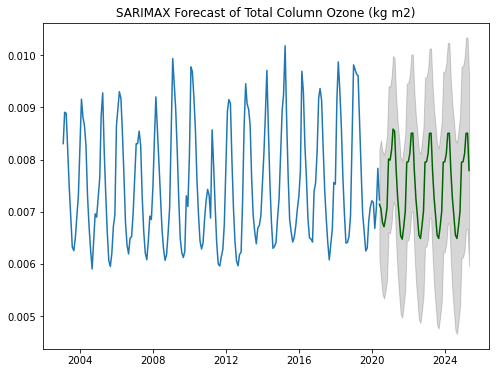
* Fitted the model and trained and tested data was put into a dataframe (converted back to scale). The RSME was 0.08529, which represents a very good accuracy score.

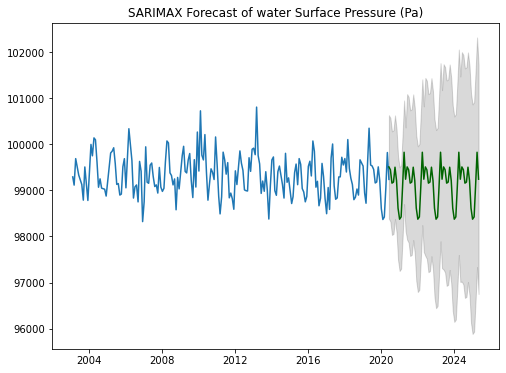
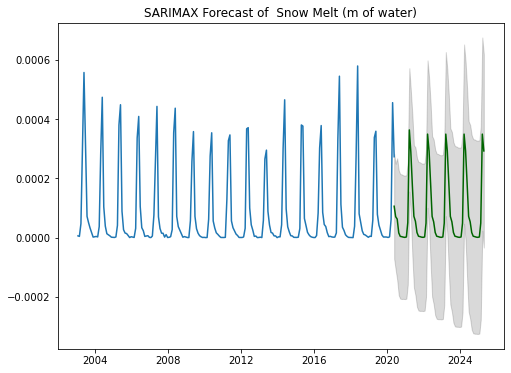


**Machine Learning - Forecasting**

After using the SARIMAX model to train the model and get a high accuracy square, we attempted to forecast the sea ice extent in the future. In order to do that:

* A univariate time-series model was applied to each of the features to estimate their future value, which are put into a dataframe





* using the predicted values of the features, we used the model to predict the values of Y (Extent):

**Website Development**