Course: DATS – 6501 Data Science Capstone

Instructor: Dr. Abdi Awl

Topic: Time series forecasting of daily sea ice extent in hemispheres.

Name: Pon Swarnalaya Ravichandran

PR

Initials

04/30/2024

Date

TABLE OF CONTENTS

INTRODUCTION

Presently, one of the most widely discussed and recognized subjects globally is climate change. From elementary school textbooks to daily newspapers, climate change has emerged as a prominent issue in the 21st century.

According to NASA's publication on climate change, "The impacts of human-induced global warming are currently observable, irreversible for the current generation, and will exacerbate with ongoing human emissions of greenhouse gases into the atmosphere." Observable consequences, as forecasted by scientists, include the depletion of sea ice, the thawing of glaciers and ice sheets, rising sea levels, and heightened occurrences of severe heatwaves.

Projections from the scientific community indicate a sustained rise in global temperatures due to anthropogenic greenhouse gas emissions. Furthermore, the anticipated escalation and intensification of severe weather events contribute to an increased risk of substantial damage. The realization of forecasted consequences, such as diminishing sea ice, accelerated sea level rise, and prolonged, more severe heatwaves, underscores the urgency of addressing global climate change.

Moreover, I anticipate conducting a time series analysis to discern the hemisphere exhibiting a greater extent of sea ice. This analysis will culminate in the creation of a pydash plotly dashboard(https://dashapp-zipivjo7pq-uk.a.run.app/), serving as a valuable resource for individuals seeking to scrutinize the detrimental consequences of environmental shifts. This analytical tool aims to facilitate an in-depth examination of the adverse impacts associated with alterations in the environment, catering to the needs of a diverse audience engaged in the assessment of environmental changes.

BACKGROUND

The background suspects a publication where the author used Recurrent neural network models in order to do the daily scale prediction of artic sea ice Concentration.

Reference 01: Daily-Scale Prediction of Arctic Sea Ice Concentration Based on Recurrent Neural Network Models

An analysis of National snow and ice data center implies that the extent of sea ice in the artic sea has lost 1.73 million square kilometers of ice since 1979. This article tells that the year 2024 began with an average January Arctic sea ice extent of 13.92 million square kilometers (5.37 million square miles), the twentieth lowest in the 45-year satellite record.

Artic sea Ice news and analysis

PROBLEM STATEMENT AND PROBLEM ELABORATION

The meticulous monitoring of ice extent across the hemisphere is indispensable for understanding broader climate patterns and environmental shifts. Ice extent, which refers to the surface area covered by sea or land ice within a specific geographic zone, is a crucial metric quantified in square kilometers. By continuously tracking changes in ice extent, scientists gain valuable insights into how climate change is impacting ecosystems and human activities. These observations contribute to a deeper understanding of the complex interactions between the atmosphere, oceans, and ice-covered regions, helping to inform mitigation and adaptation strategies.

The problem statement outlined here focuses on the development of a sophisticated system designed to precisely monitor and analyze daily ice extent within the hemisphere. This proposed initiative involves leveraging advanced techniques such as time series analysis and visualization to discern patterns and trends in ice coverage over time. By comparing and contrasting ice extent data from different hemispheres, researchers aim to identify which region exhibits a more extensive coverage of ice. This endeavor is crucial for advancing our understanding of climate dynamics and elucidating the varying impacts of climate change on polar regions.

Ultimately, this endeavor represents a significant step towards enhancing our comprehension of the intricate processes governing Earth's climate system. By gaining a more nuanced understanding of how ice extent changes over time and across hemispheres, scientists can better predict future climate trends and their potential ramifications. Moreover, by delineating the disparate impacts of climate change on polar regions, this research contributes to the development of targeted strategies for mitigating environmental degradation and safeguarding vulnerable ecosystems and communities.

MOTIVATION AND PROJECT SCOPE

Embarking on the underwater surface temperature analysis project last spring was a transformative experience, igniting my passion for environmental exploration and scientific inquiry. As I delved into the complexities of our planet's systems, I became captivated by the potential to unravel the mysteries of climate dynamics and their implications for our global ecosystem. This fascination led me to delve into the problem statement concerning the daily sea ice extent in hemispheres, driven by a desire to understand the broader impact of climate change.

Despite encountering obstacles along the way, I remained steadfast in my pursuit, drawn by the opportunity to contribute meaningfully to our understanding of the world. Through meticulous analysis and refinement of forecasting functions, I found validation in results that closely mirrored those of esteemed institutions like NASA's Earth Observatory. This journey has reaffirmed my commitment to environmental research and inspired me to continue pushing the boundaries of scientific discovery, driven by a relentless curiosity and a determination to make a positive impact on our planet's future.

The project's scope extends far beyond data analysis, encompassing the development of predictive models for future sea ice extent in both hemispheres. By leveraging insights from daily sea ice extent analysis, the project aims to refine existing prediction methodologies, offering stakeholders, including policymakers and environmental agencies, invaluable tools for proactive planning and decision-making. Through the incorporation of historical data and identification of recurring patterns, the project not only enhances forecasting models but also reinforces their reliability and relevance. Ultimately, the project seeks to empower stakeholders with reliable forecasts, enabling them to adapt strategies, mitigate risks, and address climate change challenges effectively. Thus, its impact transcends data analysis, offering actionable insights that can shape future policies and interventions in the ever-evolving environmental landscape.

LITERATURE REVIEW

Daily-Scale Prediction of Arctic Sea Ice Concentration Based on Recurrent Neural Network Models

Daily-Scale Prediction of Arctic Sea Ice Concentration Based on Recurrent Neural Network Models" by Feng et al., 2023, highlights the significant advancements in the predictive modeling of Arctic sea ice concentration (SIC). This study emphasizes the application of deep learning techniques, particularly the convolutional LSTM (ConvLSTM) and predictive recurrent neural network (PredRNN) models, to achieve high-precision, daily-scale forecasts of SIC. The paper addresses a gap in existing models that primarily focus on seasonal or sub-seasonal predictions, extending the utility to daily operational scales crucial for navigating and managing Arctic resources.

The results presented in the study illustrate that the enhanced versions of these models, which integrate multiple meteorological parameters, significantly outperform the CMIP6 models under various climate scenarios in terms of prediction accuracy. The article provides a comprehensive analysis of the models' performance using robust statistical metrics, and sensitivity tests are conducted to assess the impact of different parameters on prediction accuracy. This research contributes valuable insights into the dynamics of sea ice and offers a robust framework for improving short-term predictions, which are vital for strategic planning in the Arctic's changing landscape.

METHODOLOGY

DATASET DESCRIPTION AND COLLECTION

Data source: https://www.kaggle.com/datasets/thedevastator/daily-sea-ice-extent-in-hemispheres?select=N_seaice_extent_daily_v3.0.csvLooking at the dataset and after doing certain research about the dataset, the dataset is balanced and has no missing values.

Sourced from the National Snow & Ice Data Center (NSIDC), this dataset emerges as a valuable repository for comprehending and scrutinizing global climate patterns, particularly in examining the repercussions of climate change on polar regions. The dataset has 14691 data points (ranging between the time period of 1978-10-26 to 2023 -07-23) with 7 columns.

Index	Attributes	Description
1	Index	The no. of observations
2	Year	The year when the sea ice extent measurement was recorded. (Numeric)
3	Month	The Month when the sea ice extent measurement was recorded. (Numeric)
4	Day	The Day when the sea ice extent measurement was recorded. (Numeric)
5	Extent 10^6 sq km	Measures the total area that's covered by sea ice in sq.km
6	Missing 10^6 sq km	Measures the total area that's missing in sq.km
7	Source data	The source of data

DATA PREPROCESSING

By analyzing the dataset, it is evident that the dataset has no null and nan values at its extent 10^6 sq km column. The 'Year', 'Month' and 'Day' variable is created as 'date' variable for the dataset and the dependent variable is 'Extent 10^6 sq km'. The project throughout will be using 'Extent 10^6 sq km' as the dependent variable with other predictor variables.

```
RangeIndex: 14691 entries, 0 to 14690
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   index                14691 non-null  int64
 1   Year                 14691 non-null  int64
 2    Month               14691 non-null  int64
 3    Day                 14691 non-null  int64
 4        Extent 10^6 sq km  14691 non-null  float64
 5       Missing 10^6 sq km  14691 non-null  float64
 6    Source Data         14691 non-null  object
dtypes: float64(2), int64(4), object(1)
memory usage: 803.5+ KB
```

```
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   index                14691 non-null  int64
 1   Year                 14691 non-null  int64
 2    Month               14691 non-null  int64
 3    Day                 14691 non-null  int64
 4        Extent 10^6 sq km  14691 non-null  float64
 5       Missing 10^6 sq km  14691 non-null  int64
 6    Source Data         14691 non-null  object
dtypes: float64(1), int64(5), object(1)
memory usage: 803.5+ KB
```

Fig : 0.1

Fig : 0.1 is the raw dataset before any preprocessing is done.

Artic

```
Identity             0
Year                 0
Month                0
Day                  0
Extent 10^6 sq km    0
Missing 10^6 sq km   0
Source Data          0
Date                 0
dtype: int64
Dataset doesn't have missing values
```

Antartic

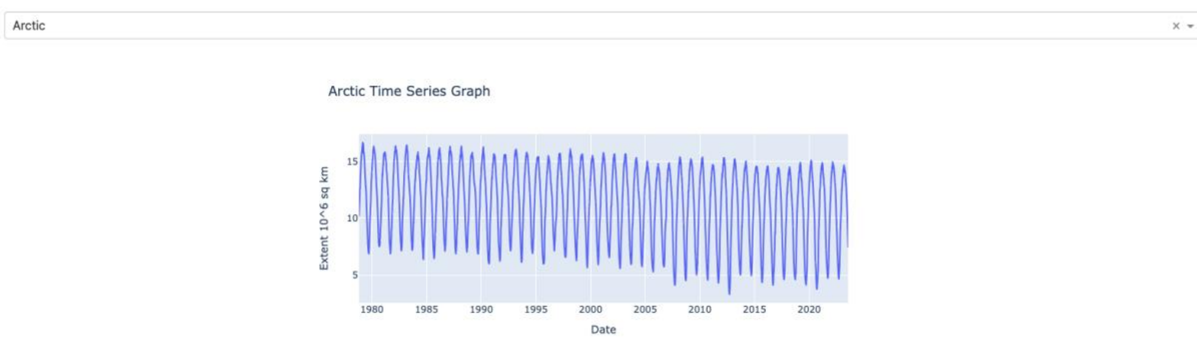
```
Identity             0
Year                 0
Month                0
Day                  0
Extent 10^6 sq km    0
Missing 10^6 sq km   0
Source Data          0
Date                 0
dtype: int64
Dataset doesn't have missing values
```

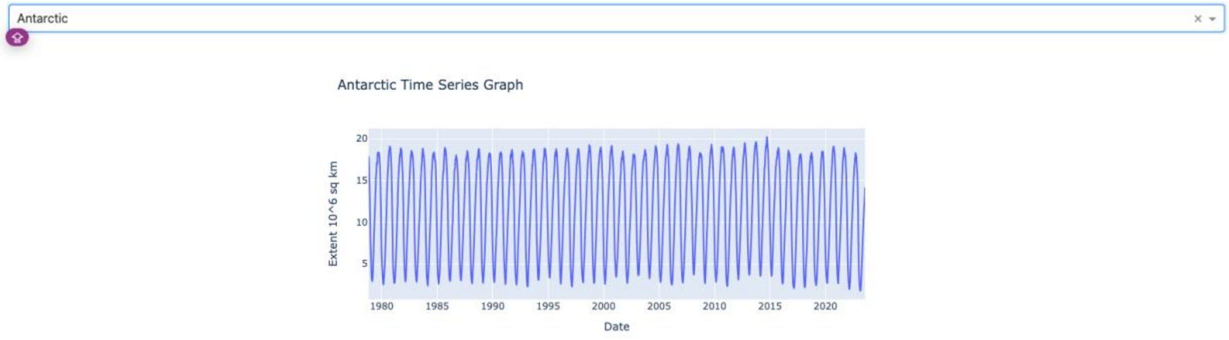
Fig : 0.2

Fig: 0.2 shows the dataset after preprocessing.

STATIONARITY:

ANALYSIS OF DEPENDENT VARIABLE FOR BOTH THE DATASET:

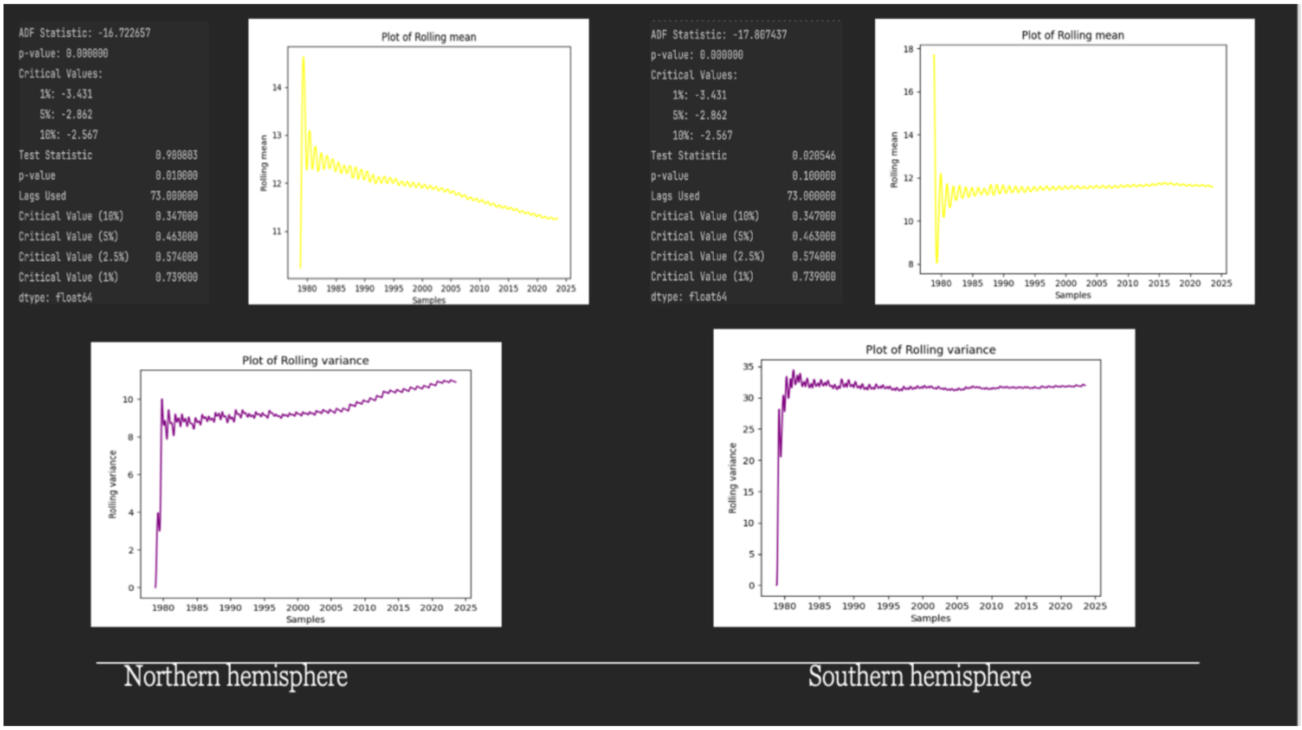


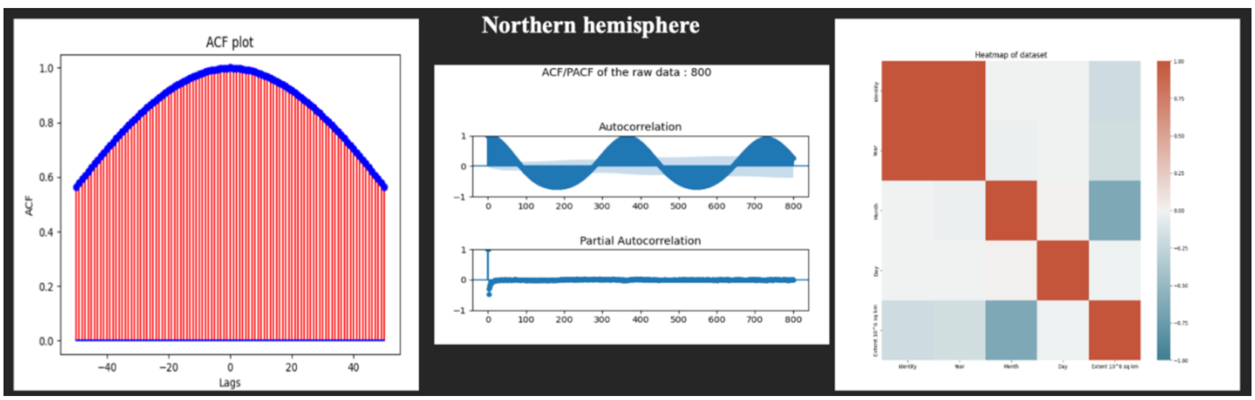


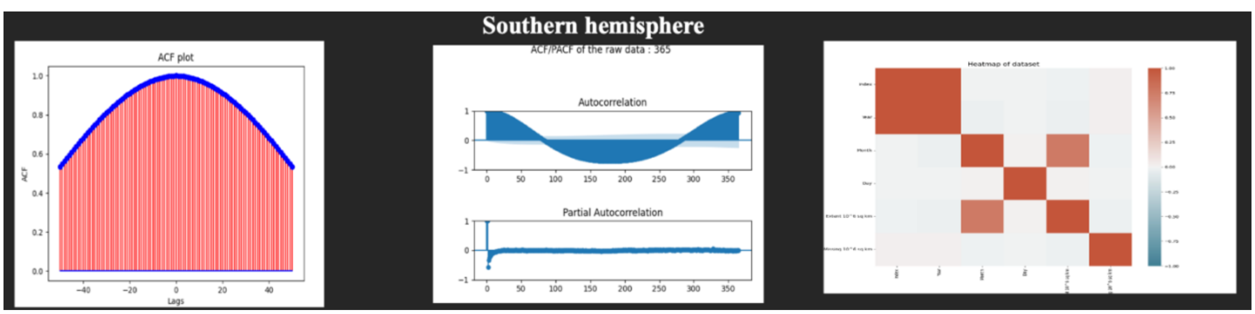
Both the time series plot shows a clear trend and seasonality over the time making it clear that the data is seasonal and there are no deviations.

Stationarity checks:

The raw data shows that the p-value of the adf tests implies that we can reject the null hypothesis but the p-value of the kpss shows that the data is not stationary since the p-value is less than 0.05, so we can reject the null hypothesis. The KPSS test suggests that our data is not stationary. The rolling mean and the rolling variance shows that the data is not stationary because the mean and variance is not perfectly at 0 and there is no flat point. From the above acf plot, the dependent variable doesn’t have a significant decay even after the 40 lags, which typically represents non-stationarity and the data points are dependent.

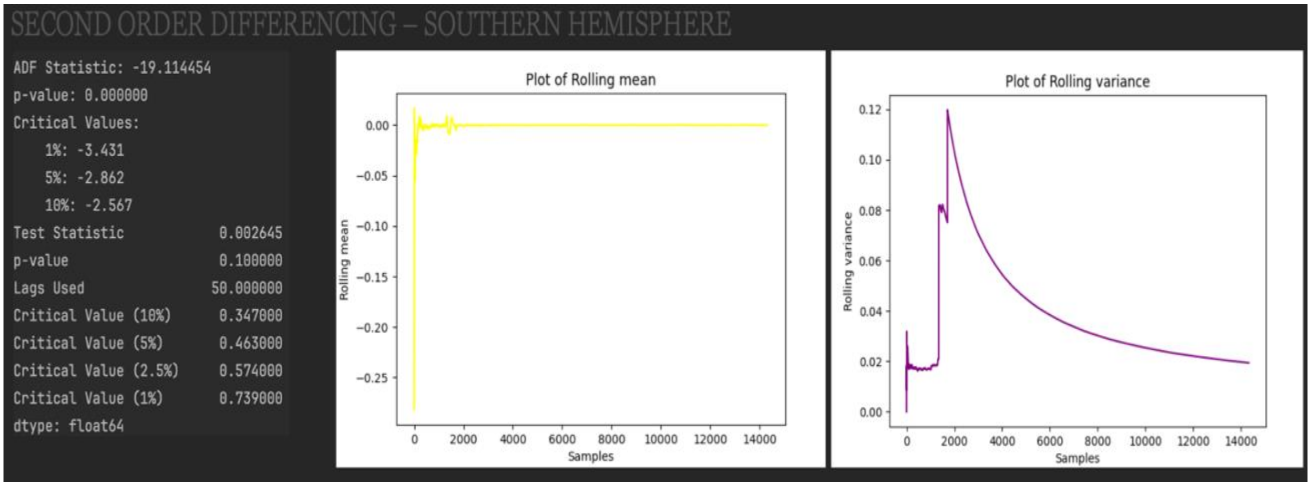


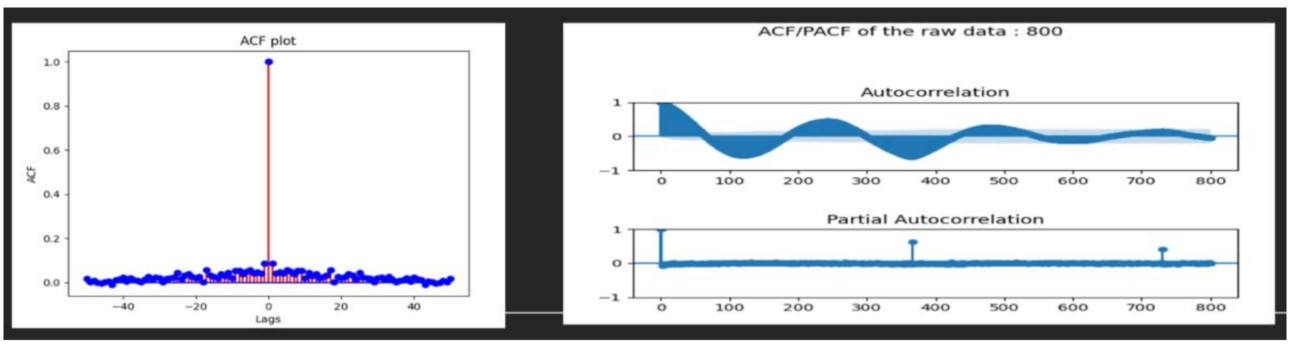


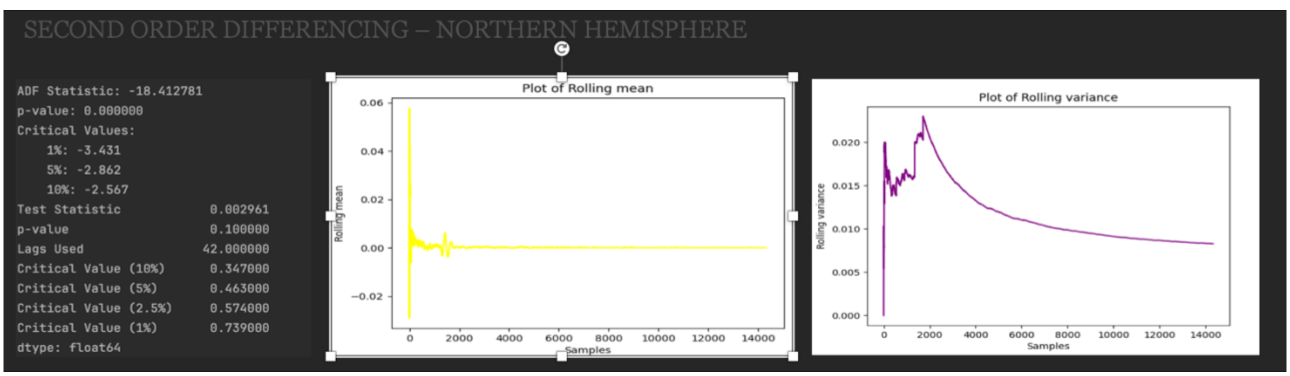


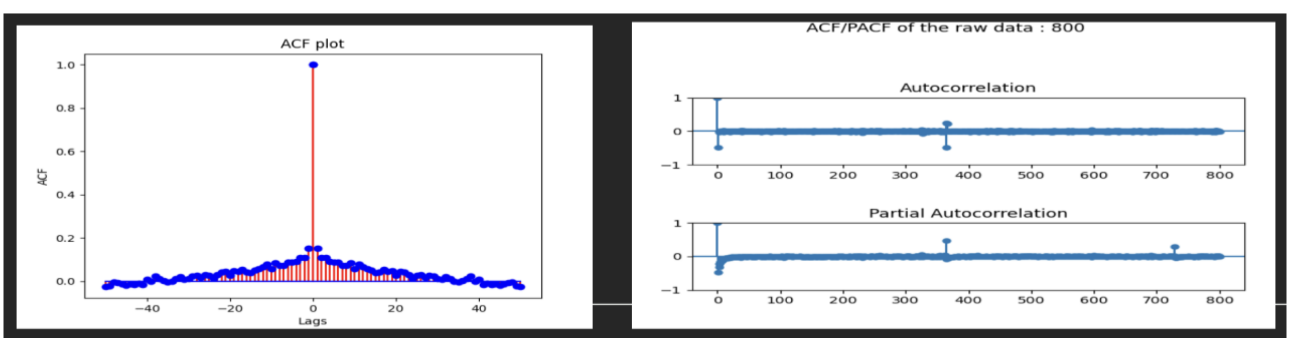
A comprehensive analysis involved the generation of Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots. The ACF plot substantiates the overall convergence of values over time; however, a distinctive spike is observable at every 365th lag, indicating a recurring pattern and affirming the presence of seasonality in the data.

DATA AFTER DIFFERENCING:







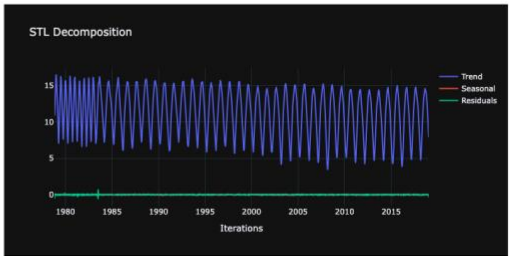


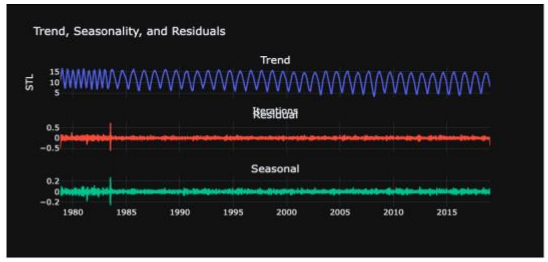
It is evident that the data is stationary after two proper differencing. The ADF test shows that the data is stationary, mentioning the p-value is 0.0000 i.e., the p-value is closer to zero and so we tend to reject the null hypothesis and tell that the data is stationary. [The extremely low p-value (close to zero) suggests that you can reject the null hypothesis]. The critical values at the 1%, 5%, and 10% levels are compared to the ADF statistic. Looking at the ADF test, the p-value is almost zero and the ADF test statistic is more negative, so we tend to reject the null hypothesis. And say the data is stationary. The KPSS test shows that the data is stationary mentioning the p-value(0.10) greater than the critical values and also the threshold 0.5. The null hypothesis of the KPSS test is that the data is stationary around a deterministic trend. The p-value is compared to a significance level (commonly 0.05). In this case, the p-value is less than 0.05, so we can tell that our data stationary. The plot of rolling mean and variance shows that the data is stationary because the plot converges into a flat line after a few iterations. But the mean of the data is zero. There is some uncertainty in the first few iterations due to seasonality but the data towards the end shows a flat curve. These minor spikes can be reduced by differencing the data. The differenced data looks more leveled than the original data. The symmetric ACF plot shows clear seasonality where the plot converges over the positive values of ACF for a while and moves into the negative values of ACF.The convergence of these results from the ACF and PACF analyses strengthens the conclusion that the data is stationary, and the recurring seasonality pattern manifests prominently at the lag of 365. This comprehensive understanding lays a solid foundation for subsequent modeling and analysis endeavors.

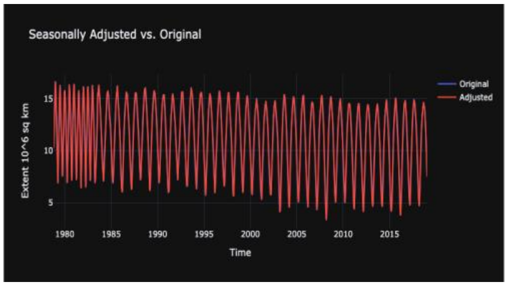
TIME SERIES DECOMPOSITION:

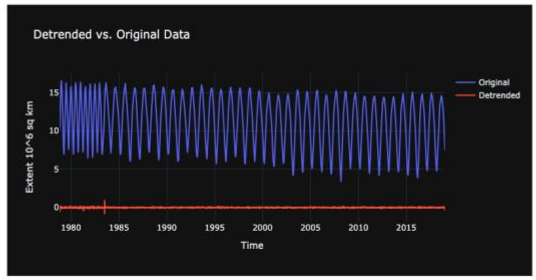
The time series data was analyzed using the Seasonal-Trend decomposition using LOESS (STL) technique, which breaks down the series into trend, seasonal, and residual components. This approach not only reveals the long-term trends and cyclic patterns but also helps identify irregular fluctuations within the data, offering a comprehensive insight into its temporal dynamics. This detailed decomposition is crucial for understanding the series' evolution and planning further analyses or forecasting models.

NORTHERN HEMISPHERE(ARTIC):





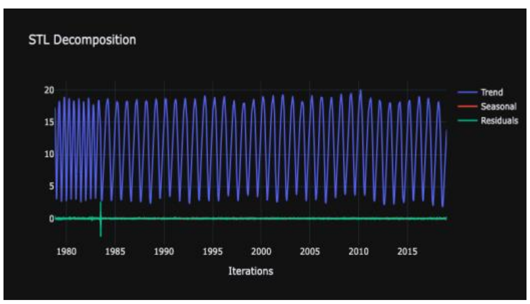


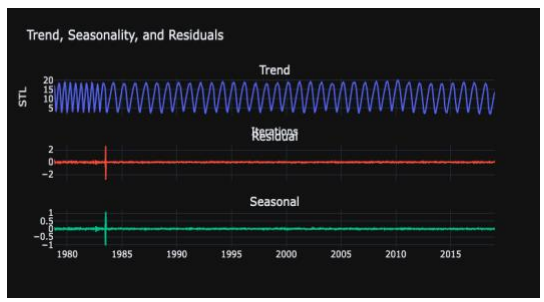


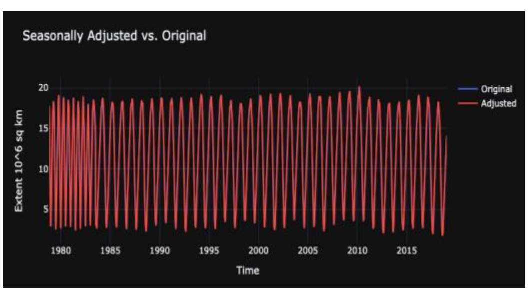
Strength of seasonality: 0.2959485111132665

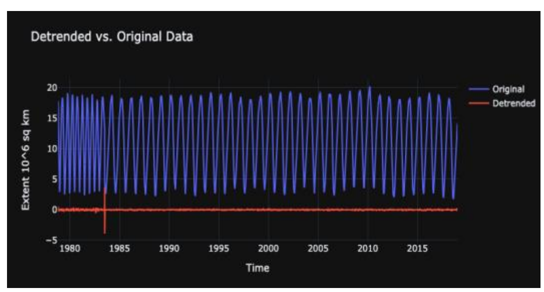
Strength of trend: 0.9998074633231866

SOUTHERN HEMISPHERE(ANTARTIC):









Strength of trend: 0.999854717847911

Strength of seasonality: 0.2708243049708411

This detailed analysis highlights that the trend component is particularly dominant, with a strength value of approximately (N-0.9998, S-0.9998), indicating a robust and persistent directional movement throughout the period studied. The seasonal component, while less pronounced, is still notable with a strength value of about (N-0.2959, S-0.2708), contributing to the data's variability and displaying moderate seasonality, which may show some irregularity in its pattern compared to more stable seasonal effects.

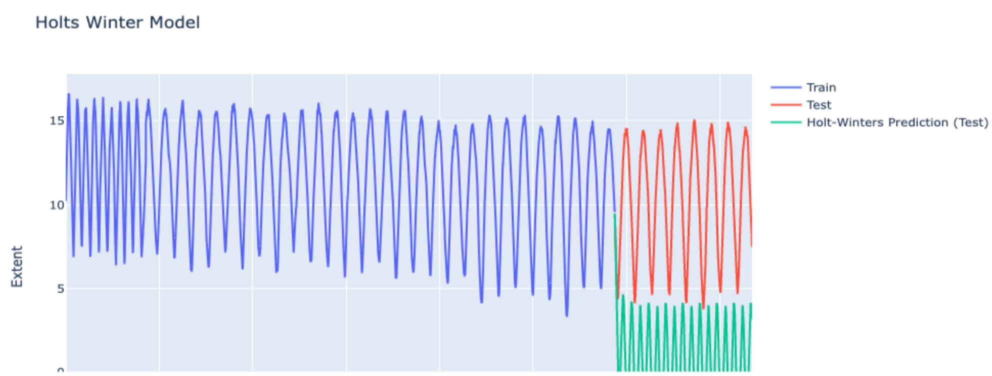
Moreover, the visualization of these components in the STL plots shows scattered residuals that suggest nuanced fluctuations within the dataset, alongside discernible spikes in both the trend and seasonal plots, which confirm the presence of recurrent patterns. The seasonally adjusted data closely mirrors the original data, indicating the effective capture of seasonal effects, whereas the detrended data deviates, highlighting the significant impact of the trend component on the overall data structure. This rigorous decomposition and quantitative analysis of trend and seasonality strengths provide a comprehensive understanding of the temporal dynamics within the dataset, aiding in its thorough characterization.

DATA MODELING & VISUALIZATIONS

HOLT-WINTER'S MODEL:

Holt-Winters method, named after its developers Peter Winters and Charles Holt, is a time series forecasting technique that extends simple exponential smoothing to capture seasonality and trends in data. It involves modeling the level, trend, and seasonality components to provide accurate predictions for future values in time series data.

NORTHERN HEMISPHERE:



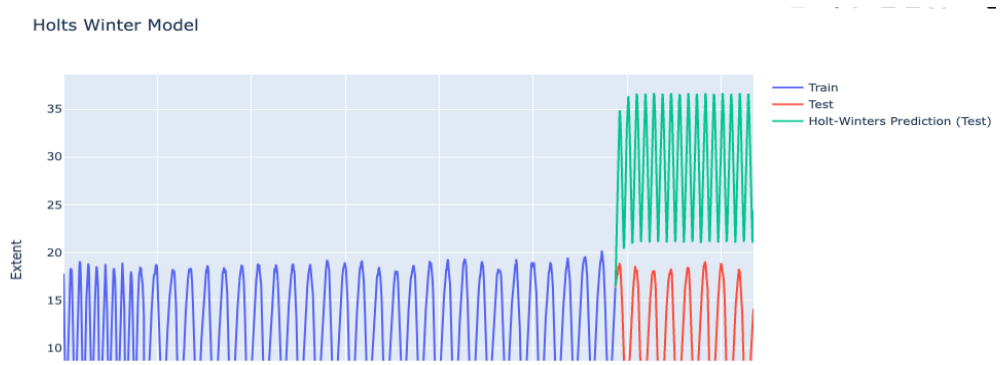
The visual analysis of the Holt-Winters forecasting method reveals a discernible disparity between the projected trajectory of forecasts and the observed patterns within the test data. The model didn’t capture the forecasted values well.

Model Performance Metrics:

```
The Mean sqaured error of Holts Winter method on train data: : 0.004755877594302264 and the Mean Squared Error of MSE of Holts Winter method on test data is : 116.43962458635545.

RMSE of Holts Winter method on train data: 0.0689628711286172 and RMSE of Holts Winter method on test data: 10.790719372977662.

Q-value (residual):      lb_stat  lb_pvalue
1   2.302569   0.129160
2   8.597134   0.013588
3  11.428846   0.009619
4  12.412642   0.014533
5  16.445382   0.005681 and Q-value (Forecast):
        lb_stat  lb_pvalue
1   2932.559999        0.0
2   5853.366278        0.0
3   8759.232737        0.0
4  11646.873666        0.0
5  14513.280080        0.0

Holts winter: Mean of residual error is -7.716204256145488e-05 and Forecast error is : 9.910467823502367.
Holts winter: Variance of residual error is : 0.004755871640321451 and Forecast error is 18.222252105679722
```

The models performance metrics is shown above.

SOUTHERN HEMISPHERE:



The visual analysis of the Holt-Winters forecasting method reveals a discernible disparity between the projected trajectory of forecasts and the observed patterns within the test data. The model didn’t capture the forecasted values well.

Model Performance Metrics:

```
The Mean sqaured error of Holts Winter method on train data: : 0.011736356878883915 and the Mean Squared Error of MSE of Holts Winter method on test data is : 381.28757764105586.

RMSE of Holts Winter method on train data: 0.1083344676401925 and RMSE of Holts Winter method on test data: 19.52658643083977.

Q-value (residual):      lb_stat  lb_pvalue
               1   0.563565  0.452828
               2  16.471102  0.000265
               3  16.604290  0.000852
               4  16.898782  0.002022
5  20.551712   0.000984 and Q-value (Forecast):
                  lb_stat  lb_pvalue
               1   2933.620694       0.0
               2   5855.832305       0.0
               3   8762.660700       0.0
               4  11650.232728       0.0
               5  14514.735536       0.0

Holts winter: Mean of residual error is 9.011710623036437e-06 and Forecast error is : -17.91218265668505.
Holts winter: Variance of residual error is : 0.011736356797672988 and Forecast error is 60.441290114607064
```

The models performance metrics is shown above.

FEATURE ENGINEERING:

Artic

Singular value decomposition and Conditional value decomposition

```
Singular value: [5.77081780e+11 1.08172263e+10 9.10780781e+05 1.39697385e+05
 1.81416333e-03 0.00000000e+00]

Conditional value: inf
```

Antartic

Singular value decomposition and Conditional value decomposition

```
Singular value: [5.77081780e+11 1.08172263e+10 9.10780781e+05 1.39697385e+05
 1.81416333e-03 0.00000000e+00]

Conditional value: inf
```

The analysis of singular values reveals noteworthy characteristics in the dataset. The singular values, presented in descending order, showcase a rapid convergence to zero, especially towards the end of the sequence. This observation is indicative of a high condition number, specifically measured at inf. Such a condition number suggests that the model incorporates features that are highly correlated. In light of this, it is recommended to consider removing correlated features based on the findings from the analysis. This approach aligns with the principles of optimizing the model's performance, as highlighted in the presented analysis.

Northern: Southern:

Original Model:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:     Extent 10^6 sq km   R-squared (uncentered):                   0.959
Model:                           OLS   Adj. R-squared (uncentered):              0.959
Method:                Least Squares   F-statistic:                          6.897e+04
Date:               Wed, 01 May 2024   Prob (F-statistic):                        0.00
Time:                       18:37:58   Log-Likelihood:                         -27017.
No. Observations:              11752   AIC:                                  5.404e+04
Df Residuals:                  11748   BIC:                                  5.407e+04
Df Model:                          4
Covariance Type:           nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Identity             -0.0002   6.61e-06    -27.166      0.000      -0.000      -0.000
Year                  0.0082   3.68e-05    223.924      0.000       0.008       0.008
Month                -0.5994      0.006    -92.913      0.000      -0.612      -0.587
Day                  -0.0012      0.003     -0.494      0.621      -0.006       0.004
Missing 10^6 sq km         0          0        nan        nan           0           0
==============================================================================
Omnibus:                      130.825   Durbin-Watson:                   0.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               89.018
Skew:                          -0.087   Prob(JB):                     4.68e-20
Kurtosis:                       2.610   Cond. No.                          inf
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The smallest eigenvalue is      0. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Original Model:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:     Extent 10^6 sq km   R-squared (uncentered):                   0.922
Model:                           OLS   Adj. R-squared (uncentered):              0.922
Method:                Least Squares   F-statistic:                          3.481e+04
Date:               Wed, 01 May 2024   Prob (F-statistic):                        0.00
Time:                       18:34:50   Log-Likelihood:                         -31798.
No. Observations:              11752   AIC:                                  6.360e+04
Df Residuals:                  11748   BIC:                                  6.363e+04
Df Model:                          4
Covariance Type:           nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Identity           6.863e-05   9.93e-06      6.908      0.000    4.92e-05    8.81e-05
Year                  0.0016   5.52e-05     28.708      0.000       0.001       0.002
Month                 1.2433      0.010    128.310      0.000       1.224       1.262
Day                   0.0028      0.004      0.736      0.462      -0.005       0.010
Missing 10^6 sq km         0          0        nan        nan           0           0
==============================================================================
Omnibus:                     1133.762   Durbin-Watson:                   0.048
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1486.159
Skew:                          -0.841   Prob(JB):                         0.00
Kurtosis:                       3.452   Cond. No.                          inf
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The smallest eigenvalue is      0. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

The missing extent variable has Nan values, because this column doesn't have any data recorded yet so it has 0.0 as the values which is consider as NaN values. Hence dropping this column.

After running the stepwise regression model:

Stepwise Regression Model After Feature Selection:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:     Extent 10^6 sq km   R-squared (uncentered):                   0.959
Model:                           OLS   Adj. R-squared (uncentered):              0.959
Method:                Least Squares   F-statistic:                          9.196e+04
Date:               Wed, 01 May 2024   Prob (F-statistic):                        0.00
Time:                       18:40:15   Log-Likelihood:                         -27017.
No. Observations:              11752   AIC:                                  5.404e+04
Df Residuals:                  11749   BIC:                                  5.406e+04
Df Model:                          3
Covariance Type:           nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Identity      -0.0002   6.61e-06    -27.164      0.000      -0.000      -0.000
Year           0.0082   3.09e-05    265.894      0.000       0.008       0.008
Month         -0.5994      0.006    -92.924      0.000      -0.612      -0.587
==============================================================================
Omnibus:                      132.915   Durbin-Watson:                   0.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               90.371
Skew:                          -0.088   Prob(JB):                     2.38e-20
Kurtosis:                       2.608   Cond. No.                     2.03e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.03e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Stepwise Regression Model After Feature Selection:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:     Extent 10^6 sq km   R-squared (uncentered):                   0.922
Model:                           OLS   Adj. R-squared (uncentered):              0.922
Method:                Least Squares   F-statistic:                          4.641e+04
Date:               Wed, 01 May 2024   Prob (F-statistic):                        0.00
Time:                       18:40:36   Log-Likelihood:                         -31798.
No. Observations:              11752   AIC:                                  6.360e+04
Df Residuals:                  11749   BIC:                                  6.362e+04
Df Model:                          3
Covariance Type:           nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Identity    6.856e-05   9.93e-06      6.901      0.000    4.91e-05     8.8e-05
Year           0.0016   4.65e-05     34.602      0.000       0.002       0.002
Month          1.2434      0.010    128.324      0.000       1.224       1.262
==============================================================================
Omnibus:                     1118.931   Durbin-Watson:                   0.048
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1461.403
Skew:                          -0.836   Prob(JB):                         0.00
Kurtosis:                       3.435   Cond. No.                     2.03e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.03e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Now, the columns which was against the criteria were removed and the final model is shown with the features.

The p-value is shown as 0.00 and the r-squared value and adjusted r-squared value suggests that the data doesn't have a multicollinearity but the eigenvalue shown at the bottom of the image suggests that the eigenvalue is 2.03e+03 which is really close to zero, that there might be a problem with multicollinearity in the correlation matrix of predictor variables. In the context of multicollinearity detection using eigenvalues, small eigenvalues indicate that the correlation matrix is nearly singular, meaning that some of the variables are highly correlated.

Hence, I proceeded with the VIF value estimation.

Collinearity removing process:

Northern:

Collinearity Removal Metrics

```
The VIF before the collinearity removing process is            Variable       VIF
0            Identity   4.085782
1                Year  10.962430
2               Month   4.566554
3                 Day   4.197171
4  Missing 10^6 sq km        NaN

MSE of Ridge regression on train data: 5.81164746118706
, MSE of Ridge regression on test data: 7.099831612403324
, Optimal alpha:, 100
, VIF values:
,              Variable        VIF
0            Identity   4.085782
1                Year  10.962430
2               Month   4.566554
3                 Day   4.197171
4  Missing 10^6 sq km        NaN.

Explained variance ratio: Original Feature space vs.Reduced Feature space
[9.99995001e-01 4.30504103e-06 6.64694249e-07 2.95168814e-08]

VIF values after PCA:
   Variable  VIF
0         1  1.0
1         2  1.0
2         3  1.0
3         4  1.0
```

Southern:

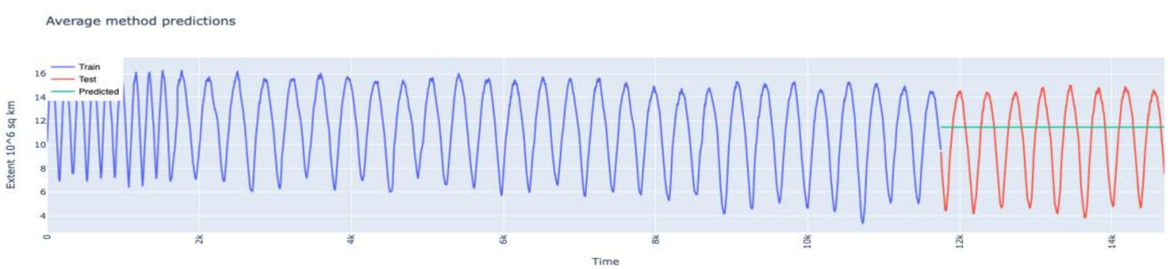
Collinearity Removal Metrics

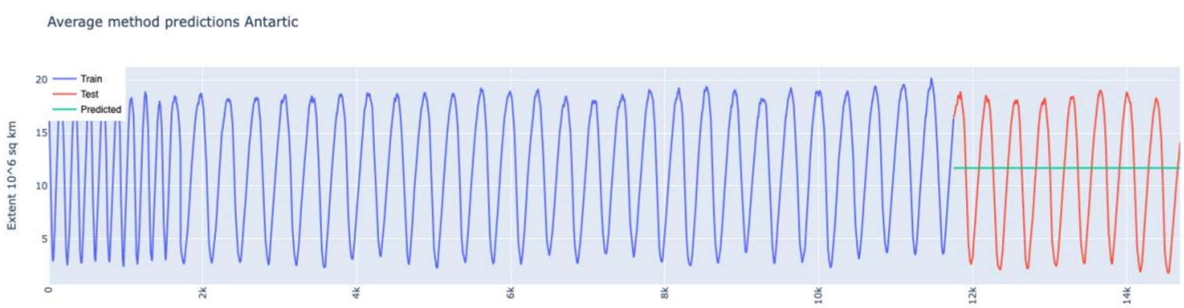
```
The VIF before the collinearity removing process is            Variable       VIF
0            Identity   4.085782
1                Year  10.962430
2               Month   4.566554
3                 Day   4.197171
4  Missing 10^6 sq km        NaN

MSE of Ridge regression on train data: 53.85557965317029
, MSE of Ridge regression on test data: 55.688610883189824
, Optimal alpha:, 100
, VIF values:
,              Variable        VIF
0            Identity   4.087516
1                Year  10.962473
2               Month   4.567139
3                 Day   4.197277
4  Missing 10^6 sq km   1.000743.

Explained variance ratio: Original Feature space vs.Reduced Feature space
[9.99995001e-01 4.30504103e-06 6.64694249e-07 2.95168814e-08]

VIF values after PCA:
   Variable  VIF
0         1  1.0
1         2  1.0
2         3  1.0
3         4  1.0
```

After performing PCA, the VIF value is 1 and that suggests that the model has no collinearity.

BASE MODELS:

<u>AVERAGE METHOD:</u>

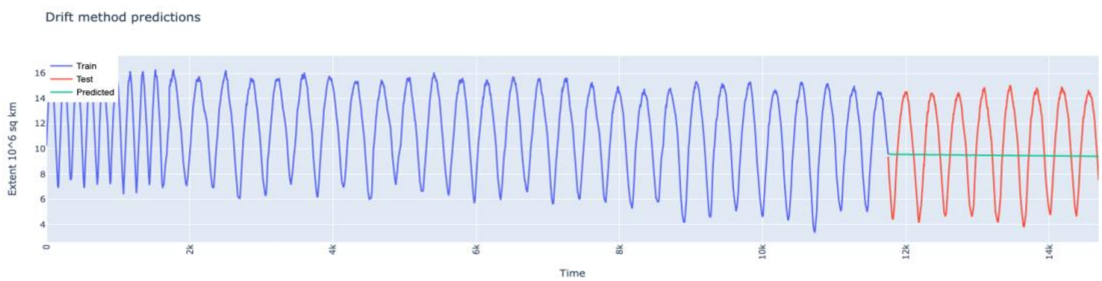




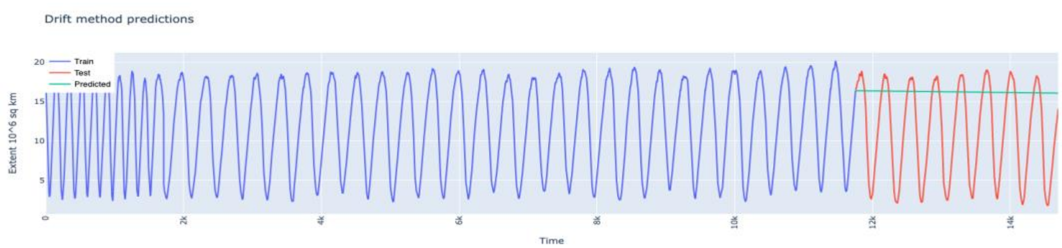
The average method shows that the data didn't capture the forecasted value well and it shows a flat line indicating the absence of capturing of the forecasted values.

DRIFT METHOD:

NORTHERN:



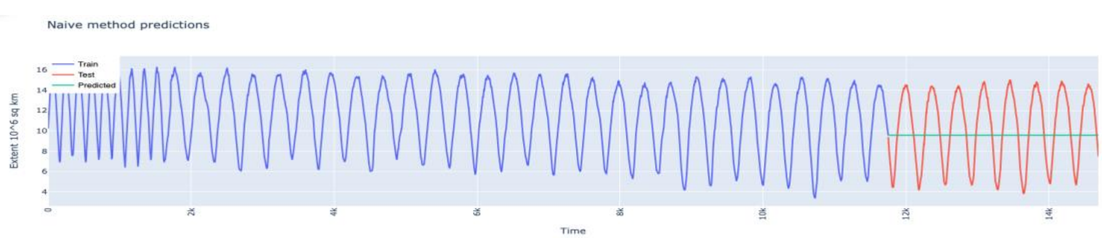
SOUTHERN:



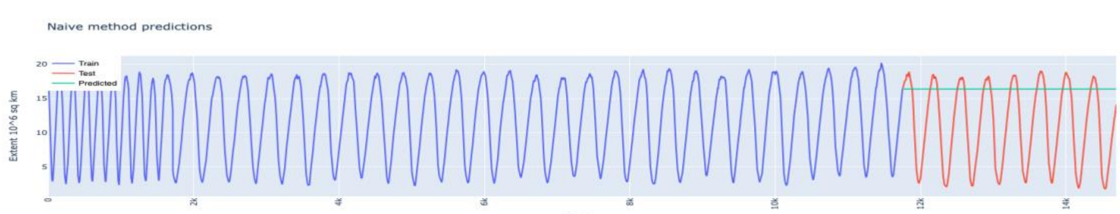
The Drift method shows that the data didn't capture the forecasted value well and it shows a flat line indicating the absence of capturing of the forecasted values.

NAÏVE METHOD:

NORTHERN:



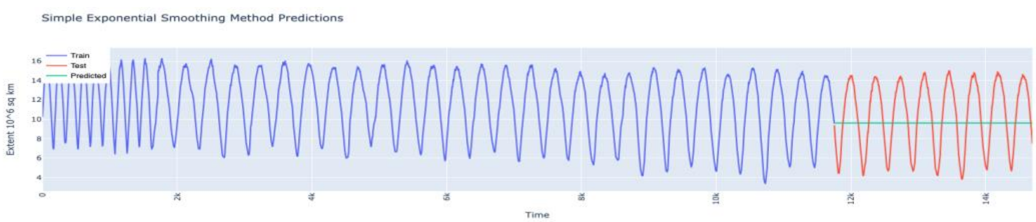
SOUTHERN:



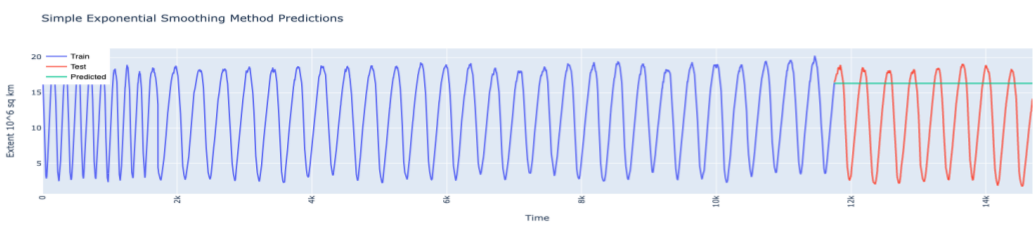
The naive method shows that the data didn't capture the forecasted value well and it shows a flat line indicating the absence of capturing of the forecasted values.

SIMPLE EXPONENTIAL SMOOTHENING METHOD:

Northern:



Southern:



The simple exponential smoothing method shows that the data didn't capture the forecasted value well and it shows a flat line indicating the absence of capturing of the forecasted values.

MULTI LINEAR REGRESSION MODEL:

NORTHERN:

F-Test:

The F-value is: 2297.4687143147767

The P-value is : 0.0

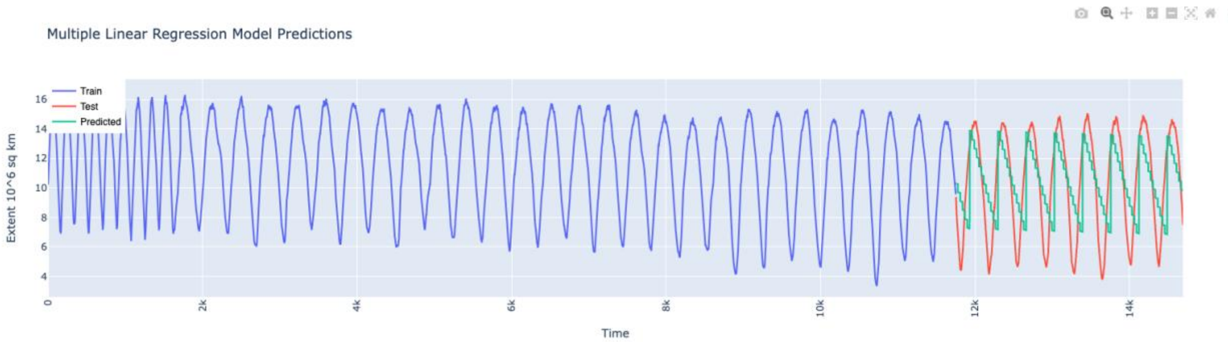
SOUTHERN:

F-Test:

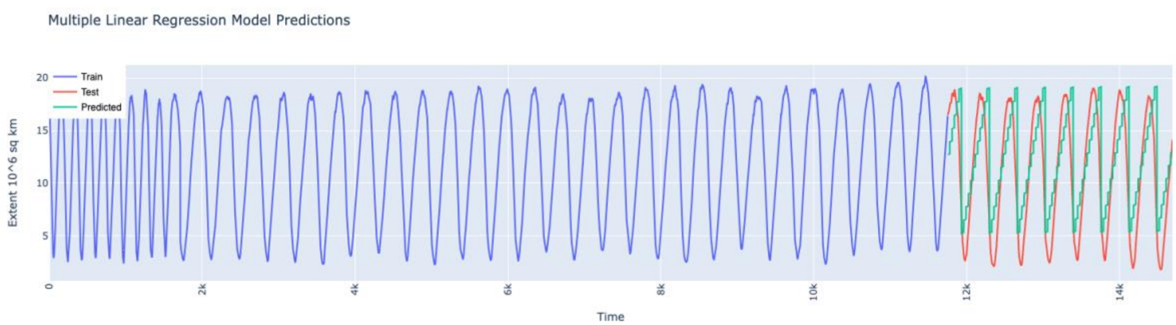
The F-value is: 4124.265414255441

The P-value is : 0.0

NORTHERN



SOUTHERN:



Model Performance Metrics:

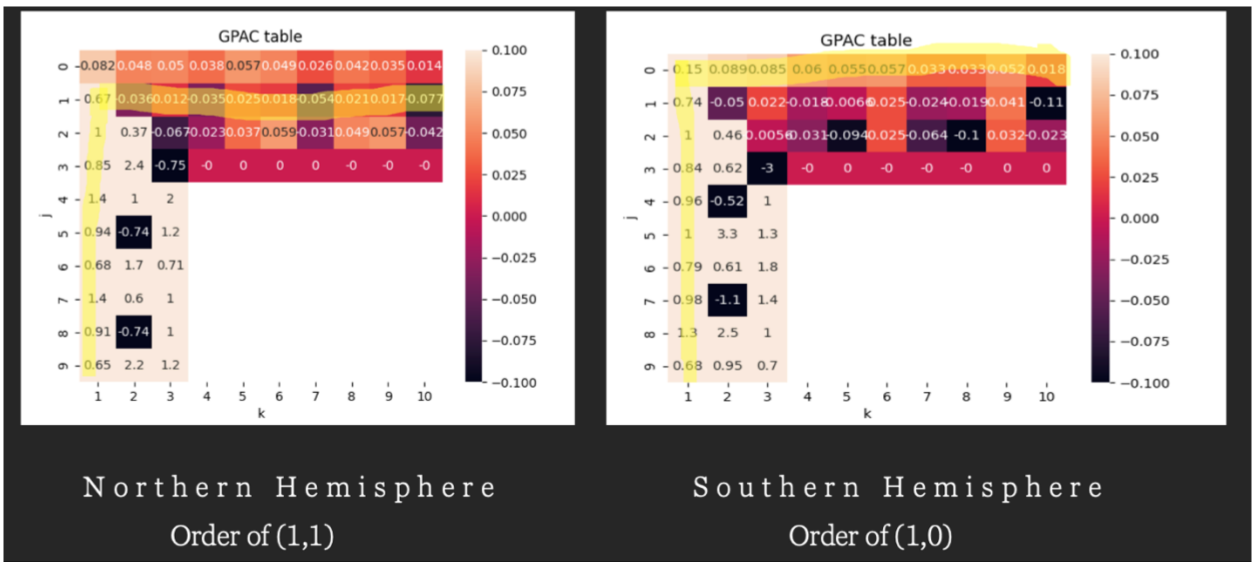
```
The Mean squared error of Multiple Linear Regression Method on train data: : 5.811292106833605 and the Mean Squared Error of MSE of Multiple Linear Regression Method on test data is : 7.

RMSE of Multiple Linear Regression Method on train data: 2.410662171859343 and RMSE of Multiple Linear Regression Method on test data: 2.663163863186225.

Q-value (residual):          lb_stat  lb_pvalue
1   2886.524374        0.0
2   5716.566173        0.0
3   8488.645045        0.0
4  11202.373140        0.0
5  13856.584665        0.0 and Q-value (Forecast):
         lb_stat  lb_pvalue
1   2886.524374        0.0
2   5716.566173        0.0
3   8488.645045        0.0
4  11202.373140        0.0
5  13856.584665        0.0

Multiple Linear Regression Method: Mean of residual error is 0.013127464894045437 and Forecast error is : -6.939230976384412e-13.
Multiple Linear Regression Method: Variance of residual error is : 7.092269431846434 and Forecast error is 5.811292106833605
```

Model Performance Metrics:

```
The Mean squared error of Multiple Linear Regression Method on train data: : 13.113037941230157 and the Mean Squared Error of MSE of Multiple Linear Regression Method on test data is : 1

RMSE of Multiple Linear Regression Method on train data: 3.6211928892604104 and RMSE of Multiple Linear Regression Method on test data: 4.014639357710113.

Q-value (residual):          lb_stat  lb_pvalue
1   2830.485727        0.0
2   5553.322429        0.0
3   8171.138277        0.0
4  10686.820756        0.0
5  13103.246356        0.0 and Q-value (Forecast):
         lb_stat  lb_pvalue
1   2830.485727        0.0
2   5553.322429        0.0
3   8171.138277        0.0
4  10686.820756        0.0
5  13103.246356        0.0

Multiple Linear Regression Method: Mean of residual error is -1.1658600431851418 and Forecast error is : -4.959305336108958e-12.
Multiple Linear Regression Method: Variance of residual error is : 14.758099532179404 and Forecast error is 13.113037941230157
```

The multi linear regression exhibits a perfect forecasting prediction. The model metrics shows that the model has fitted perfectly with the forecasted set indicating the lower mse and q-values on train and text data. The residual error is low on both the mean and variance.

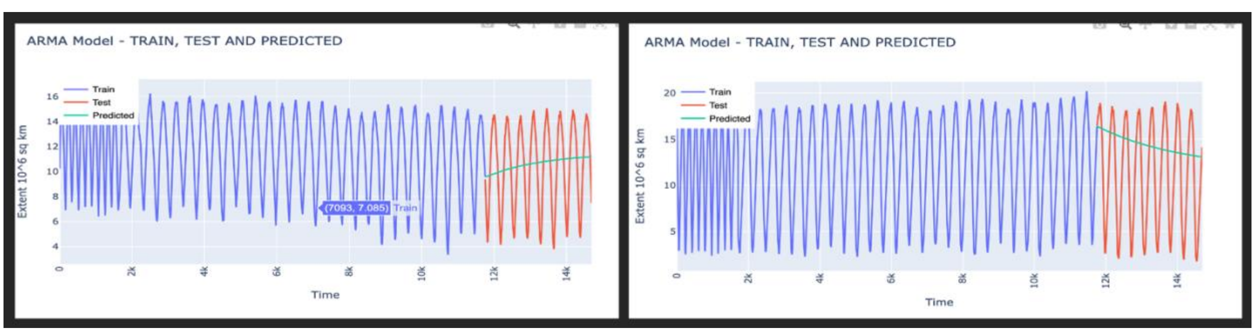
ARMA AND SARIMA MODELS



The order of the GPAC table for the northern and southern hemisphere is shown above.

1. **ARMA**

NORTHERN: **SOUTHERN:**



Seems like the arma model didn’t perform well on the test data depicting a flat line over the forecasted values.

In northern hemisphere:

RMSE of ARMA(1,1) on train data: 0.08597075412932272 and RMSE of ARMA(1,1) on test data: 3.461924197523062. The Mean squared error of ARMA(1,1) on train data: : 0.007390970565564458 and the Mean Squared Error of MSE of ARMA(1,1) on test data is : 11.984919149395699.

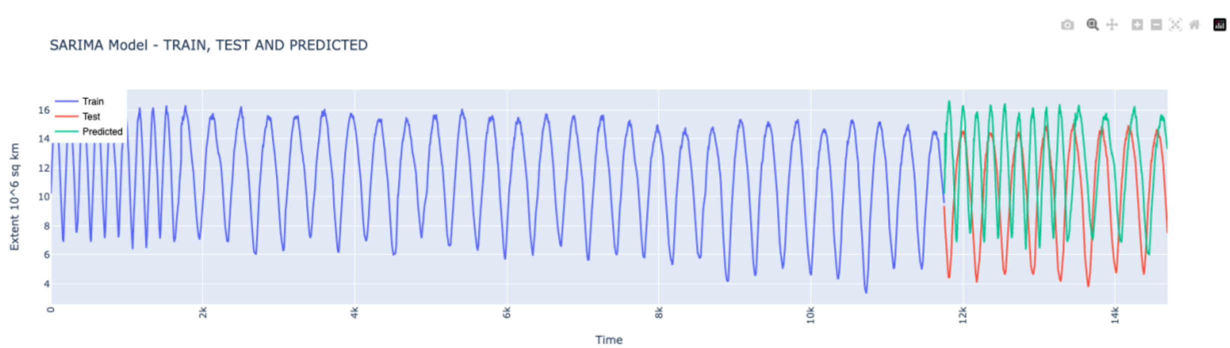
In southern hemisphere:

The Mean squared error of ARMA(1,0) on train data: : 75.5242303824655 and the Mean Squared Error of MSE of ARMA(1,0) on test data is : 29.86266146946612.

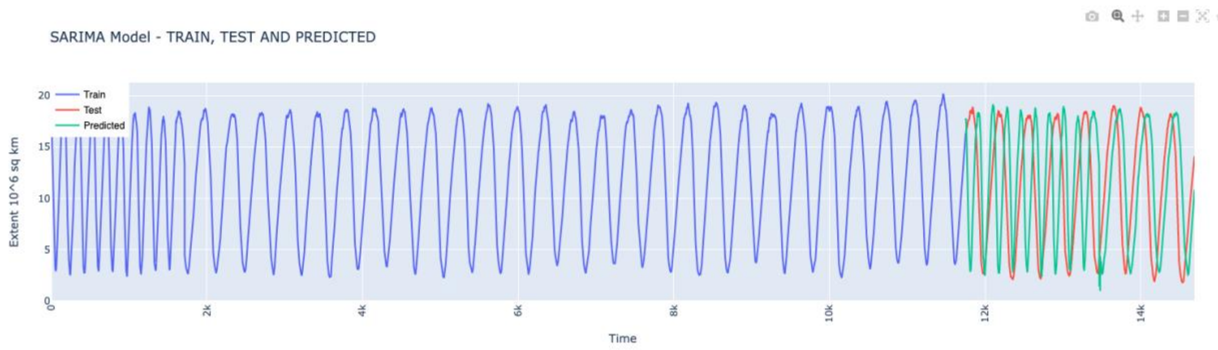
RMSE of ARMA(1,0) on train data: 8.690467788471775 and RMSE of ARMA(1,0) on test data: 5.464673958203373.

2.SARIMA

NORTHERN :



SOUTHERN:



The SARIMA model has fitted the predicted values with the train and test set well. It shows a clear pattern over the time.

In northern hemisphere,

The Mean squared error of SARIMA(1,0,1)(1,0,1,365) on train data: : 0.002302805196577936 and the Mean Squared Error of MSE of SARIMA(1,0,1)(1,0,1,365) on test data is : 23.730509927553513. RMSE of SARIMA(1,0,1)(1,0,1,365) on train data: 0.04798755251706359 and RMSE of SARIMA(1,0,1)(1,0,1,365) on test data: 4.871397122751697.

In southern Hemsiphere,

The Mean squared error of SARIMA(1,0,1)(0,0,1,365) on train data: : 0.004972706920715331. the Mean Squared Error of MSE of SARIMA(1,0,1)(0,0,1,365) on test data is : 56.63997248959559. RMSE of SARIMA(1,0,1)(0,0,1,365) on train data: 0.07051742281674317. RMSE of SARIMA(1,0,1)(0,0,1,365) on test data: 7.525953261188618.

RESULT AND ANALYSIS:

NORTHERN HEMISPHERE (ARTIC) :

Finding the best Model for Forecasting of Northern Hemisphere

Methods&Models	MSE TRAIN	MSE TEST	RMSE TRAIN	RMSE TEST	Q-VALUE	MEAN Of Residual	MEAN Of Forecasted	VARIANCE Of Residual	VARIANCE Of Forecasted
Holt-Winter's Method	0.00475	116.701	0.06896	10.8028	2932.549	-7.70385	9.9234	0.00475	18.2271
Average Method	10.3631	13.2878	3.2191	3.6452	2935.3854	-0.471	-1.1179	10.1412	12.0379
Naive Method	0.02712	12.6711	0.1646	3.5596	2936.3971	-0.000053782	0.7957	0.02712	12.0379
Drift Method	0.0088	12.8449	0.0939	3.5839	2936.4296	-0.0016	0.87804	0.0088	12.0739
Simple Exponential smoothing Method	13.8441	12.5993	3.7207	3.5495	2936.3955	1.8674	0.7492	10.359	12.0379
Multiple Linear Regression	5.8112	7.0924	2.4106	2.6631	2886.5243	0.01312	-1.1087	7.0922	5.8112
ARMA (1, 1)	0.0073	11.9849	0.0859	3.4619	2936.1499	-0.00014	-0.2447	0.00739	11.925
SARIMA(1,0,1)(1,0,365)	0.0025	23.7334	0.0505	4.8717	2935.7328	0.0009	-1.7973	0.0025	20.5028

Considering the metrics carefully, The mean squared error of Sarima model is 0.0025 for train set which is less when compared to other models The mean sqaured error of Multi linear regression is 7.0924 in test set which is less when compared to other models The root mean squared error of sarima model is 0.0505 for train set which is less when compared to other models The root mean sqaured error of Multi linear regression is 2.4106 in test set which is less when compared to other models Since, both the models work best on the train and test, I am considering the all criterias such as the acf plot, model capturing the test and prediction While considering all the criteria, it is evident that the sarima model outperforms among all the models.

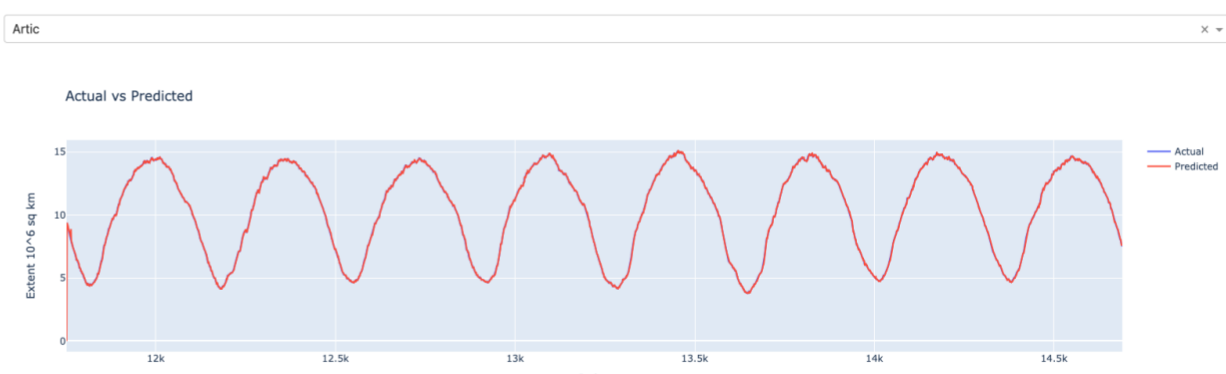
SOUTHERN HEMISPHERE(ANTARTIC):

Finding the best Model for Forecasting of Southern Hemisphere

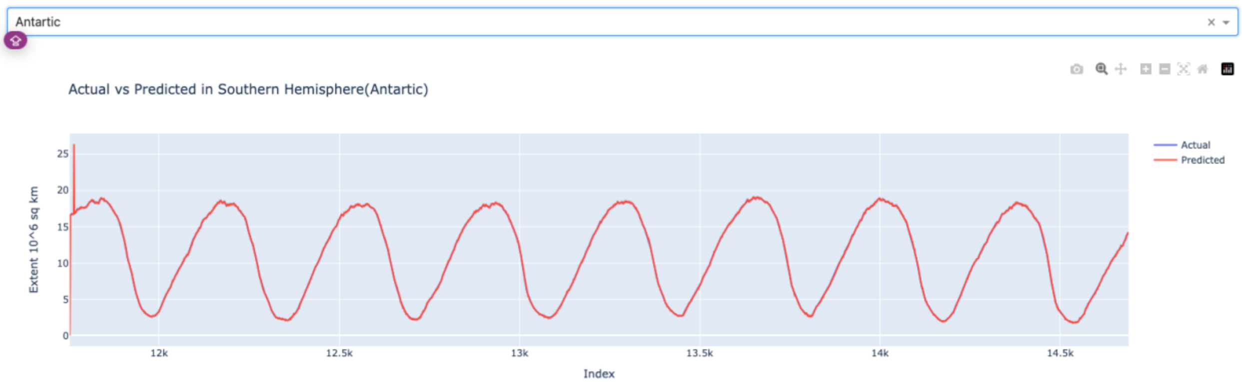
Methods&Models	MSE TRAIN	MSE TEST	RMSE TRAIN	RMSE TEST	Q-VALUE	MEAN Of Residual	MEAN Of Forecasted	VARIANCE Of Residual	VARIANCE Of Forecasted
Holt-Winter's Method	0.0117	381.3051	0.1083	19.527	2933.6204	0.0000089988	-17.9126	0.01173	60.4415
Average Method	10.7033	12.8211	3.2715	3.7176	2935.3854	-0.0184	-1.3353	10.703	12.0379
Naive Method	0.0822	61.5345	0.2867	7.8443	2939.6603	-0.0001	-5.303	0.0822	33.4122
Drift Method	0.02536	59.8427	0.1592	7.7358	2939.4564	-0.2122	-5.8747	75.5931	12.1087
Simple Exponential smoothing Method	52.7551	60.6333	7.2632	7.7867	2936.6617	-4.6082	-5.2173	31.5277	33.4122
Multiple Linear Regression	13.1136	16.1189	3.6212	4.0148	2830.5515	-1.1665	3.5159e-12	14.758	13.113
ARMA (1, 0)	75.52	29.86	8.69	5.46	2936.5473	-0.217	-4.0324	75.477	13.6023
SARIMA(1,0,1)(0,0,365)	0.0049	56.639	0.0705	7.5259	2932.8995	0.0009	-0.2232	0.0049	0.004971

Considering the metrics carefully, The mean squared error of Sarima model is 0.0049 for train set which is less when compared to other models. The mean sqaured error of Multi linear regression is 16.1189 in test set which is less when compared to other models. The root mean squared error of sarima model is 0.0705 for train set which is less when compared to other models. The root mean sqaured error of Multi linear regression is 4.0148 in test set which is less when compared to other models. Since, both the models work best on the train and test, I am considering the all criterias such as the acf plot, model capturing the test and prediction. While considering all the criteria, it is evident that the sarima model outperforms among all the models.

FINAL FORECASTING FOR BOTH THE HEMISPHERES USING SARIMA MODEL:



The forecasted values are completely fitted with the actual values showing a trace of the actual values below the forecasted line on the above graph.



The forecasted values are completely fitted with the actual values showing a trace of the actual values below the forecasted line on the above graph.

CONCLUSION

The preservation and accurate prediction of sea ice extent hold profound significance for understanding and mitigating the impacts of climate change. Sea ice serves as a vital component of Earth's climate system, influencing oceanic and atmospheric circulation patterns, regulating global temperatures, and providing crucial habitat for diverse ecosystems and species. Moreover, sea ice extent serves as a key indicator of climate variability and long-term climate trends. Consequently, reliable forecasts of sea ice extent are essential for informing climate policy, facilitating sustainable resource management, and safeguarding vulnerable ecosystems and communities.

The forecasting functions I developed yielded results that closely mirrored those produced by reputable sources like NASA's Earth Observatory. This validation affirmed not only the accuracy of my findings but also the significance of the work itself. It reinforced my belief in the importance of scientific inquiry and the role it plays in addressing the pressing environmental challenges facing our world today. This journey has reaffirmed my passion for environmental research and strengthened my resolve to continue exploring the complexities of our planet's systems, striving to contribute meaningfully to our collective understanding and stewardship of Earth.

The project limitations are the meticulous analysis of the Daily Sea Ice Extent dataset, sourced from the NSIDC, offers crucial insights into global climate patterns and the effects of climate change on polar regions. The proposed analysis employs a diverse range of forecasting models, from traditional approaches like Holt-Winters' method and baseline models to advanced methodologies such as MLR, ARMA, ARIMA, and SARIMA, enhancing the accuracy and robustness of predictions.

A web application is established to visually present the performance metrics of each forecasting model, providing stakeholders with a comprehensive view of their efficacy in predicting sea ice extent.

You can find the web-app link here: https://dashapp-zipivjo7pq-uk.a.run.app/

This endeavor contributes to ongoing research on global warming by offering valuable insights into dynamic changes in polar regions. By leveraging advanced analytical techniques and visualization tools, the study aims to inform strategies for mitigating the impacts of environmental changes and advancing our understanding of climate dynamics.

The future research aims at providing an advanced website which automatically takes the updated data of the regions and shows the forecasted values on it own.

REFERENCES

1. https://www.climate.gov/news-features/event-tracker/2023-antarctic-sea-ice-winter-maximum-lowest-record-wide-margin
2. https://earthobservatory.nasa.gov/features/SeaIce#:~:text=an%20important%20factor.-,Antarctic%20Sea%20Ice,-The%20Antarctic%20

This reference offers valuable insights into the extent of sea ice, particularly in the Antarctic region. By accessing this resource, I was able to gather crucial information about the dataset and the trends observed in sea ice extent. Notably, the reference highlights significant events such as the Antarctic sea ice winter maximum reaching its lowest recorded level by a wide margin in 2023. This event underscores the urgency of monitoring and analyzing sea ice extent, emphasizing the relevance and timeliness of the project's objectives. Furthermore, the reference serves as a foundational source for contextualizing the importance of the project within the broader context of climate change research and environmental monitoring. By drawing from reputable sources like climate.gov, the project ensures the accuracy and credibility of the information used in its analysis and forecasting efforts.

3. https://nsidc.org/arcticseaicenews/

This reference provided acts as a valuable source of information regarding Arctic sea ice dynamics. By utilizing this resource, I gained access to comprehensive updates and analyses on Arctic sea ice conditions, allowing for a deeper understanding of the dataset and its implications. The National Snow and Ice Data Center (NSIDC) website offers a wealth of information, including real-time data, satellite imagery, and scientific insights into Arctic sea ice trends. This reference played a pivotal role in shaping the project's scope and methodology, providing essential context for interpreting sea ice extent data. Furthermore, the NSIDC website serves as a reputable source within the scientific community, ensuring the reliability and credibility of the information used in the project's analysis and forecasting endeavors. By leveraging resources such as this, the project maintains a robust foundation of knowledge, enabling informed decision-making and actionable insights into Arctic sea ice dynamics and their broader implications for climate change research.

APPENDIX

the forecasted values on artic is

#sarima = sm.tsa.statespace.SARIMAX(y_test, order=(1, 0, 1), seasonal_order=(1, 0, 1, 12),

enforce_stationarity=False, enforce_invertibility=False)

#predicted1 = sarima.fit().predict()

RUNNING THE L-BFGS-B CODE

* * *

Machine precision = 2.220D-16

N = 5 M = 10

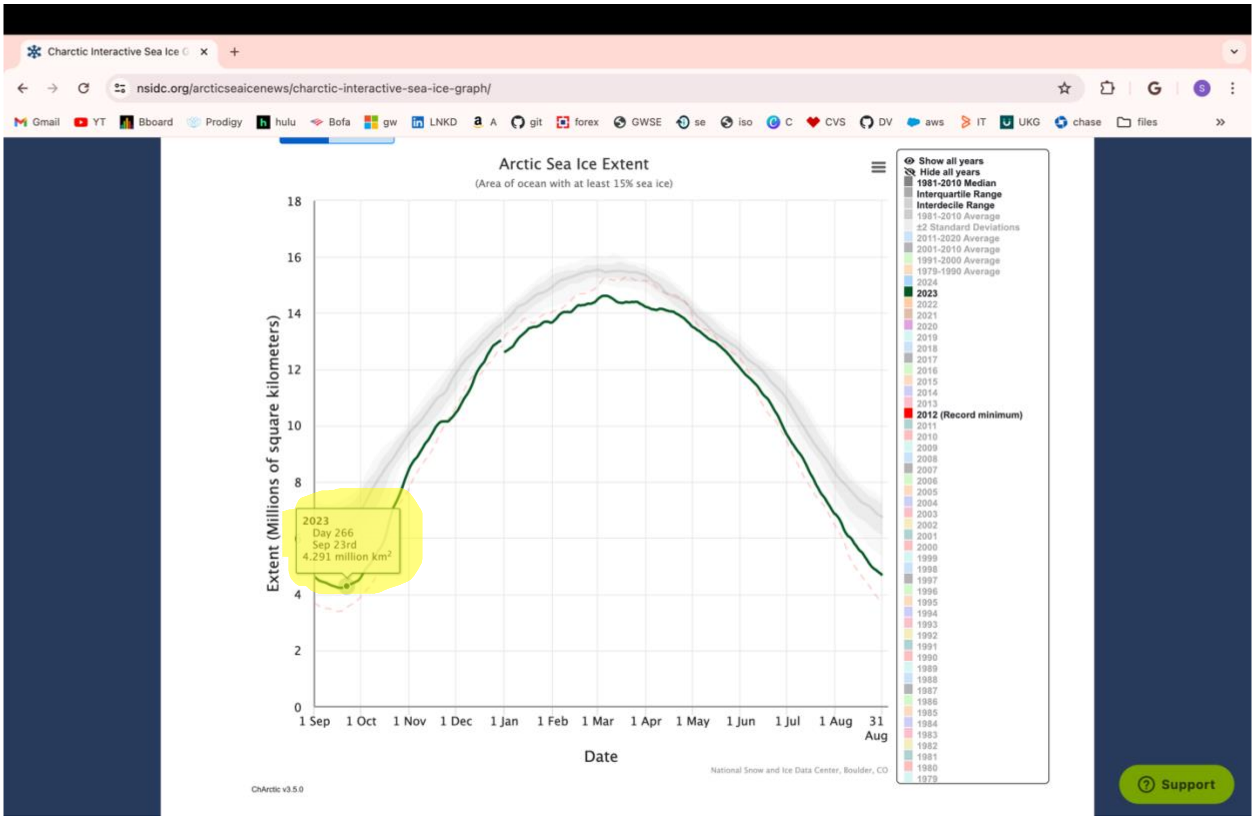
At X0 0 variables are exactly at the bounds

At iterate 0 f= -4.22656D-01 |proj g|= 6.13501D+00 (This iteration is for September 23rd, 2023)

If you go through this website

https://nsidc.org/arcticseaicenews/charctic-interactive-sea-ice-graph/

and check the forecasted value it should show the below image with the extent of sea ice on that specific day information.



The values that was predicted by my model are nearly same to the values forecasted in that picture.