# Finding Causal Factors behind Sea Level Rise, Analysis and Prediction using Machine Learning in Bay of Bengal

Project Report submitted to the Indian Institute of Technology, Kharagpur in fulfillment of the requirements for award of Degree of

**BACHELOR OF TECHNOLOGY**
in
**OCEAN ENGINEERING AND NAVAL ARCHITECTURE**

**And**

**MASTER OF TECHNOLOGY**
in
**OCEAN ENGINEERING AND NAVAL ARCHITECTURE**

by

**Mr. Amber Shekhar Shrivastava** **Roll No.: 18NA30003**

**Under the Supervision of**
**Dr. Prasad Kumar Bhaskaran**
**Professor**

**DEPARTMENT OF OCEAN ENGINEERING AND NAVAL ARCHITECTURE**
**INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR**
**(Public technical and research university established by the government of India in 1951) Kharagpur, West Bengal - 721302**

# INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

**(Public technical and research university established by the government of India in 1951)**

**Kharagpur, West Bengal - 721302**

## Ocean Engineering and Naval Architecture

## Certificate of Approval

This is to certify that the project work entitled "Finding Causal Factors behind Sea Level Rise, Analysis and Prediction using Machine Learning in Bay of Bengal" is a bonafide work carried out by Mr. Amber Shekhar Shrivastava (Roll No. 18NA30003) in partial fulfillment for the award of Master of Technology in Ocean Engineering and Naval Architecture in Indian Institute of Technology Kharagpur during the year 2022 - 2023. It is certified that all corrections/ suggestions indicated for assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

**Dr. Prasad Kumar Bhaskaran**
Supervisor and Professor
Department of Ocean Engineering and Naval Architecture,
Indian Institute of Technology, Kharagpur, West Bengal - 721302

# INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

**(Public technical and research university established by the government of India in 1951)**

**Kharagpur, West Bengal - 721302**

## Ocean Engineering and Naval Architecture

## Certificate of Examination

This is to certify that the project work entitled "Finding Causal Factors behind Sea Level Rise, Analysis and Prediction using Machine Learning in Bay of Bengal" is a bonafide work carried out by Mr. Amber Shekhar Shrivastava (Roll No. 18NA30003) in fulfillment for the award of Master of Technology in Ocean Engineering and Naval Architecture in Indian Institute of Technology Kharagpur during the year 2022 - 2023. The approval does not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn as recorded in this thesis. It only signifies the acceptance of the thesis for the purpose for which it is submitted.

External Examiner : _____

Supervisor : _____

**Dr. Prasad Kumar Bhaskaran**

Head of the Department : _____

**Dr. Vishwanath Nagarajan**

# DECLARATION

I do hereby declare that the Project Report entitled **"**Finding Causal Factors behind Sea Level Rise, Analysis and Prediction using Machine Learning in Bay of Bengal**"** is a record of an original work done by me under the guidance of Dr. Prasad Kumar Bhaskaran, Professor in the Department of Ocean Engineering and Naval Architecture, Indian Institute of Technology, Kharagpur. This report is submitted by me in the partial fulfillment of the requirements for the award of degree of Master of Technology in Ocean Engineering and Naval Architecture to Indian Institute of Technology, Kharagpur in the month of April, 2023. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Yours sincerely,

**Name : Amber Shekhar Shrivastava** **(Sign. with date)**
**Roll no : 18NA30003**
**Degree programme : Ocean Engineering and Naval Architecture**

# ACKNOWLEDGEMENT

I would like to express my sincere thanks to the supervisor, Dr. Prasad Kumar Bhaskaran, Professor, Ocean Engineering and Naval Architecture for his morale boosting, meticulous guidance, co-operation and supervision throughout this project work.

I would like to convey my sincere thanks to the Management of our institute for providing me required infrastructure within campus.

I would like to thank our Head of Department Prof. Vishwanath Nagarajan for his guidance and support and I would also like to thank both teaching and non-teaching staff members of the Ocean Engineering and Naval Architecture for their kind cooperation and all sorts of help to complete this project work successfully.

I would also like to thank all of my juniors, classmates and my friends for their valuable suggestions required to complete my project work in time.

Last but not the least I would like to thank my parents for always staying beside me and encouraging me all the time.

**Date :**

**Place : Indian Institute of Technology, Kharagpur**

**Mr. Amber Shekhar Shrivastava** **Roll no. : 18NA30003**

# ABSTRACT

Sea level estimation is important for many operations pertaining to diverse fields and operations such as building coastal structures, harbor operations and for the common people at large as well. Through this report, I have analyzed sea level change in the Bay of Bengal region at 21.25°N, 87.75°E, and have tried to analyze the level based on the tools available. There are many factors that influence sea level but it is now commonly known that the sea level is rising at a much faster pace because of Climate change.

When looking at all the evidence, there is a large scientific consensus that humans are the leading cause of climate change. In their latest report, the Intergovernmental Panel on Climate Change (IPCC) stated unequivocally that human activity is the cause of global warming. Natural climate cycles can change the temperature of Earth, but the changes we are seeing are happening at a scale and speed that natural cycles cannot explain. These cycles affect the global temperature for years, or sometimes just months, not the 100 years that we have observed. Human activity, such as burning fossil fuels and changing how we use the land, is the leading cause of climate change.

Sea level rise poses a serious threat to coastal life around the world. Some of the consequences include increased intensity of storm surges, flooding, and damage to coastal areas. In many cases, this is where large population centers are located, in addition to fragile wildlife habitats. Therefore, it is necessary to analyze and find out the causal factors that cause sea level to rise so as to assess what is coming next and to plan accordingly and mitigate the disaster that might unfold if we fail to study the change properly.

# CONTENTS


| Title | | Page No. |
|---|---|---|
| **TITLE PAGE** | | **I** |
| **CERTIFICATE** | | **II** |
| **DECLARATION** | | **III** |
| **ACKNOWLEDGEMENT** | | **IV** |
| **ABSTRACT** | | **V** |
| **CONTENTS** | | |
| **LIST OF FIGURES** | | |
| **CHAPTER 1** | **INTRODUCTION** | **10** |
| **CHAPTER 2** | **LITERATURE REVIEW** | **11** |
| **CHAPTER 3** | **PROBLEM DEFINITION** | **12-13** |
| **CHAPTER 4** | **DATASET USED FOR THE STUDY** | **14-19** |
| | 4.1 For Predicting Sea Level | 14 |
| | 4.2 For Causal Factor and Analysis | 15 - 19 |
| | 4.2.1 Downloading data | |
| | 4.2.2. Compiling the data | |
| **CHAPTER 5** | **METHODOLOGY** | **20 - 31** |
| | 5.1 For Predicting Sea Level | 20 - 23 |
| | 5.1.1 Checking for Stationarity | |
| | 5.1.2 ARIMA Model | |
| | 5.2 For Analysis and finding the Causal Factors | 24 - 31 |
| | 5.2.1 Regression Analysis | |
| | 5.2.2 XGBoost | |
| | 5.2.3 Causal Inference | |

| | | |
|---|---|---|
| **CHAPTER 6** | **RESULTS AND SUMMARY** | **32 - 43** |
| | 6.1 Prediction Results | 32 - 35 |
| | 6.2 For Finding Causal Factor behind the rise | 36 - 42 |
| | 6.2.1 Regression Analysis Results | |
| | 6.2.2 XGBoost Results | |
| | 6.2.3 Granger Causality Results | |
| | 6.2 Conclusion and Summary | 43 |
| **CHAPTER 7** | **FUTURE WORK** | **44** |
| **CHAPTER 8** | **REFERENCES** | **45** |

# LIST OF FIGURES


| S. No. | Title | Page No. |
|---|---|---|
| | **Chapter 3: Problem Definition** | |
| 1. | 21.25°N,87.75°E with respect to India | 12 |
| 2. | Close up location of 21.25°N,87.75°E wrt to Haldia Port | 12 |
| 3. | Sea level(in centimeters) for the year 1993 and 2021 respectively | 13 |
| | **Chapter 4: Dataset Used for the Study** | |
| 5. | Sea level per day from 1993 onwards | 14 |
| 4. | Correlation heatmap of the raw data | 16 |
| 5. | Correlation heatmap generated after dropping intercorrelated variables | 17 |
| | **Chapter 5: Methodology** | |
| 6. | Rolling mean and rolling std deviation | 20 |
| 7. | Test statistics for Dickey Fuller Test | 21 |
| 8. | OLS Regression Results | 25 |
| 9. | Structural Causal Model | 31 |
| | **Chapter 6: Result and Summary** | |
| 10. | Original data along with its decomposed components | 32 |
| 11. | Trendline and the best fit line of the same | 33 |
| 12. | Seasonality graph for the first 1700 days | 33 |
| 13. | ACF Plot | 34 |
| 14. | PACF Plot | 34 |
| 15. | Original along with the ARIMA fitted values | 34 |
| 16. | Predicted and original sea level (month wise) from the year 1993 onwards | 35 |
| 17. | Original and predicted sea level along with the 95% confidence level | 35 |
| 18. | Regression analysis result | 36 |
| 19. | OLS Regression result continued | 37 |
| 20. | Feature importance graph by variables | 38 |
| 21. | Feature importance graph after hyper parameter tuning of the model | 39 |
| 22. | Sea level average and the river discharge plotted together wrt number of days | 40 |
| 23. | Sea level average and the potenial_vorticity at 1hPa plotted together | 40 |
| 24. | Sea level average and the sea surface temperature plotted wrt days | 41 |

# 1

# INTRODUCTION

The changing climate has affected both global and regional meteorological conditions and sea level. Generally, sea level change occurs due to a rise in water levels around Earth. According to the 4th assessment report of the Intergovernmental Panel on Climate Change, global warming could lead to a global sea level rise of almost 60 cm by 2100. Due to the diverse meteorological and climatic conditions in various places, such an increase would not be uniform. For two key causes, studying sea level variation is extremely important. First, changes in Earth's climate have a direct impact on how quickly sea levels are rising. Second, residents in coastal areas suffer significant socioeconomic effects as a result of sea level rise. From a coastal engineering point of view, sea level change may cause erosion and flooding problems. This situation poses a direct threat to individuals and coastal structures. Therefore, it is crucial for the sustainable design and operation of coastal structures to analyze and forecast changes in sea level.

In recent years, ML techniques have provided significant advances in the modeling of physical variables at both global and local scales. These techniques can help identify complex relationships between variables as well as finding relationships with its lagged value. This way, we can optimally combine multiple data towards explainability and prediction of events, for which we need consistent sea level data along with other variables that may influence sea level rise so as to make our analysis and prediction robust. There are many tidal gauge stations that measure sea level on a daily basis but very few of them provide the data to the common public. Also, the gauges provide data to a very specific location which at times may not be the one which we are searching for. Therefore, sea level data from Copernicus Climate Data Storage were retrieved from their website which provides satellite sea level data for 720 latitude and 1440 longitude points i.e 1 million points on earth for each day from the year 1993 onwards. Along with sea level there are many other indicators available such as humidity, wind speed, surface temperature etc. Given all of the above variables might give us some clue and correlation with the sea level but the anomaly or extreme events cannot be ignored as it is an important factor which influences sea level.

For this report, the focus is mainly on analyzing Sea level along with its short term and long term prediction and finding the causal factor behind the change based on the available data. Use of modern ML algorithms such as ARIMA model, XGBoost, Regression Analysis, Correlation heatmap and Granger Causality were used to achieve the above stated goals.

# 2
# LITERATURE REVIEW

In this project, sea level rise was analyzed and predicted using the 30 year sea level data from the year 1993 onwards for which a variety of resources were used as reference. Some of them are listed below.

.

“Predicting regional coastal sea level changes with Machine Learning” by Veronica Nieves, Cristina Radin and Gustau Camps which was published in the ‘Nature’ articles dated 07 April, 2021 was used as reference to know how to clean and preprocess the data and how to assess the scope of Machine Learning for predicting sea level.

“Sea Level Prediction using Machine Learning” by Rifat Tur, Erkin Tas, Ali Torabi and Ali Mehr was taken as reference to know more about the use of Machine Learning techniques such as Linear Regression and adaptive neuro-fuzzy inference system in the prediction of sea level.

Copernicus documentation was referred to understand more about NetCDF files and the variables present in each file and how to access ir from the given files.

“ARIMA Model - Complete guide to time series forecasting” by Selva Prabhakaran which was published on Aug 22, 2022 was used to know more about the functioning of the SARIMA Model.

“Interpreting ACF and PACF plots for Time series forecasting” by Leonie Monigatti which was published on ‘TowardsDataScience’ dated Aug 3, 2022 was used to understand the use cases of ACF and PACF plots and to know more about its significance.

"ARIMA Models and the Box–Jenkins Methodology" by Asteriou, Dimitros, Hall, Stephen G. was used to know more about the internal working and methods inside of the ARIMA model.

“Machine Learning in Causal Inference” by Y Zao, Y Yu, H Wang was used to know more about how to better understand Causal inference.

# 3

# PROBLEM DEFINITION

The aim of the project is to analyze and to possibly predict the sea level in the Bay of Bengal region using only the sea level data by analyzing historical and seasonal trends and patterns combined with the anomaly that comes up along the way. The below images show us the location for which the study was conducted which is 21.25°N,87.75°E.

Fig 3.1 - 21.25°N,87.75°E with respect to India

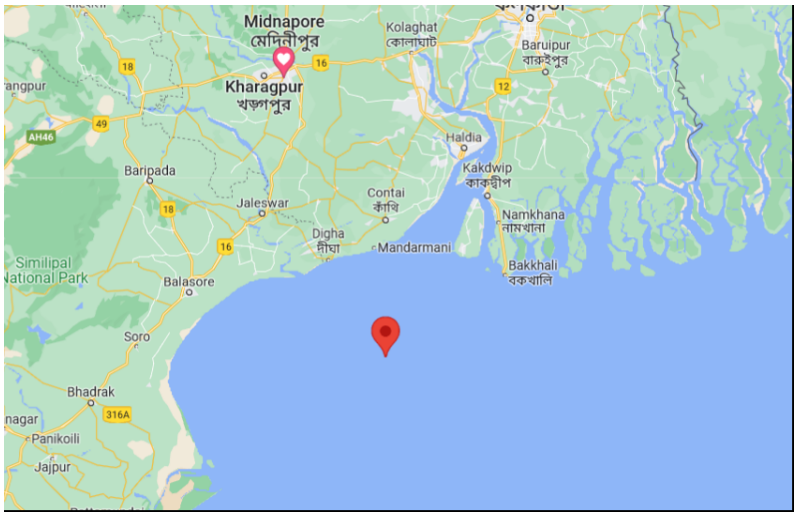


Fig 3.2 - Close up view of location of 21.25°N,87.75°E wrt to Haldia Port

The following graph shows sea level(in centimeters) for the year 1993 and 2021.

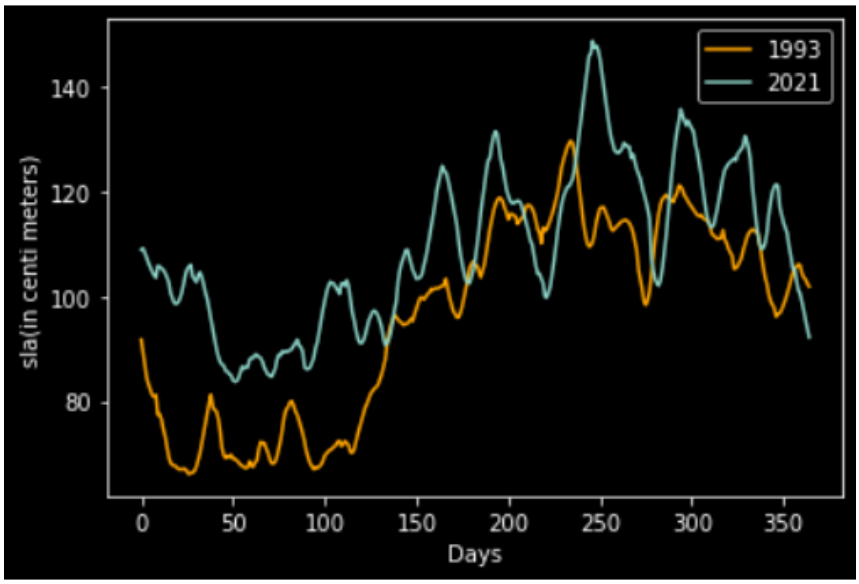


Fig 3.3 - sea level(in centimeters) for the year 1993 and 2021 respectively

We can observe from the figure that sea level in 2021 is significantly higher than it was in 1993 which is 29 years apart. The annual average sea level is rising after each passing year and thus it is important for us to investigate and analyze what are the factors that are leading to this cause. It is also equally important to pinpoint the causal factors based on the available data so as to mitigate the disaster that is unfolding before us.

Finding causal factors is an arduous task and it will require much more data, patience and hardwork to reach any conclusive point. It is thus necessary to prepare a comprehensive dataset consisting of multiple variables that can possibly have an effect on the sea level.

# 4
# DATASET USED FOR THE STUDY

## 4.1 FOR PREDICTING SEA LEVEL

The data pertaining to '*sea level anomaly'* (Sea surface height above mean sea surface computed with respect to a 20-year mean reference period 1993 to 2012) and '*absolute dynamic topography'* (Sea surface height above the geoid computed as the sum of the sea level anomaly with the mean dynamic topography) were retrieved from the copernicus website for the year 1993 to 2022 respectively for predicting the sea level rise.
Each year's data was to be downloaded through the API request, which contained 365 NetCDF files, each of the file contained the sea level data along with other data for the entire globe of that particular day in the year.

Due to the large size of the file(3 GB per file), all the data could not be processed all at once. Therefore, performed the following transformation on the data -

1. Merged all the data present in 365 files into a single file.
2. Extracted the sla(in m) and adt(in m) for the 11 latitude (i.e. 19.375 to 21.875) and 40 longitudes (i.e. 86.125 to 90.875) which were relevant to the study in the Bay of Bengal.
3. Saved the extracted data into a pickle file and deleted the rest to save space
4. Performed the above 3 steps to all the 30 files (from the year 1993 to 2022)
5. Merged the data present in all the pickle files to obtain final daily sea level data from the year 1993 onwards for the grid 19.375N,86.125E to 21.875N,90.875E.

The sea level average(sla) raw data plot

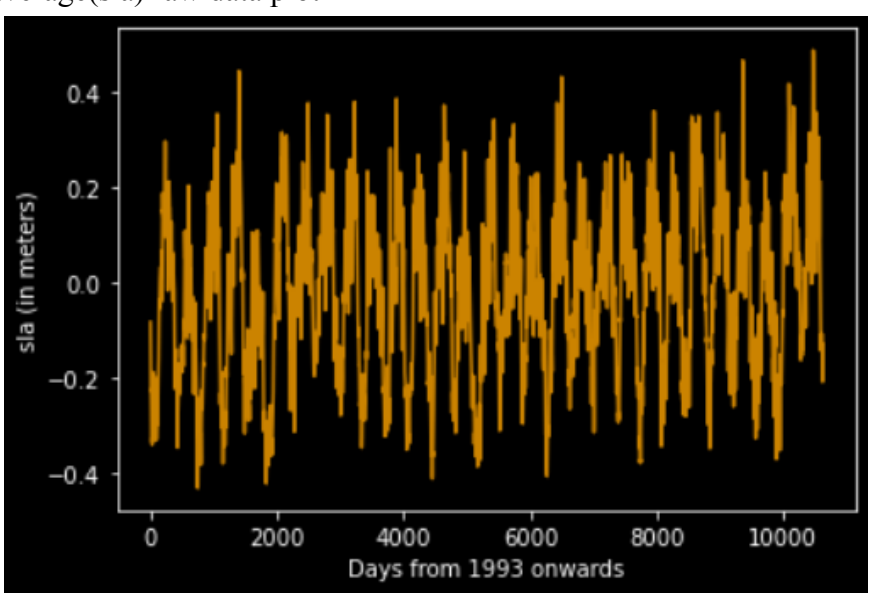


Fig 4.1.1 : Sea level per day from 1993 onwards

## 4.2 FOR ANALYSIS AND FINDING THE CAUSAL FACTORS BEHIND THE CHANGE

### 4.2.1 DOWNLOADING THE DATA

The data pertaining to '*potential_vorticity, u_wind, v_wind, w_wind, divergence, temperature'* was retrieved for the period 1990 to 2022 from the ERA5 which is the fifth generation ECMWF reanalysis for the global climate and weather for the past 8 decades. Retrieved data for the pressure levels 1hPa to 1000hPa.

The daily data penetrating to '*u100(u-component wind velocity at 100m), v100(v-component wind velocity at 100m), u10(u-component wind velocity at 10m), v10(v-component wind velocity at 10m), t2m(temperature at 2m), uvb(downward UV radiation at the surface), mbld(mean boundary layer dissipation), mpww(Mean period of wind waves), msl(Mean sea level pressure), msdwuvrf(Mean surface downward UV radiation flux), mwd(Mean wave direction), mwp(Mean wave period), pp1d( peak wave period), sst(Sea surface temperature), swh(Significant height of combined wind waves and swell), shww(Significant height of wind waves), sp(Surface pressure)'* were retrieved from the ERA5 which provides hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities for the period of 1990 to 2020.

The data pertaining to '*Surface level for River Discharge(in the last 24hrs)*' was retrieved for the period 1990 to 2022 from the Copernicus dataset which was produced by forcing a hydrological modelling chain with ERA5 meteorological reanalysis data, interpolated to the GloFAS resolution, produced at a 24-hourly timestep. Two variations of the ERA5 forcing data are used, resulting in two types of river discharge data, intermediate and consolidated. Intermediate river discharge data is produced using ERA5 Near Real Time (ERA5T) data and is updated daily, whilst consolidated data is produced using the consolidated ERA5 reanalysis and is updated monthly.
All the data above were retrieved for the region 20N,87E to 22N,89E and the required data was retrieved for our Study area.

The data pertaining to 'Average annual temperature', 'population', 'manufacturing', 'fuel imports', consumption of various items for India were retrieved from the World bank database for the year 1960 to 2020. The necessary information was gathered from the accompanying sources- World Bank, India Data, 1960-2020

### 4.2.2 COMPILING THE DATA

Since all the data used were not taken from a single source, therefore it is necessary to ensure that the taken data is correct and there is no mismatch between the length of variables. Therefore, it is necessary to clean the data and compile it so that they all can be in one place. While compiling the data, the following factors were taken into account so as to include a particular variable in the model.

1. The abundance of data throughout the required period.
2. R2 score and correlation matrix
3. Resultant wind speed and other variables that seemed relevant for our study.

If the data is monthly, interpolation is required so that the size of the variable is the same as the rest of the columns. After all the operations, a DataFrame is constructed using the pandas library having 10220 rows and 48 columns in total.

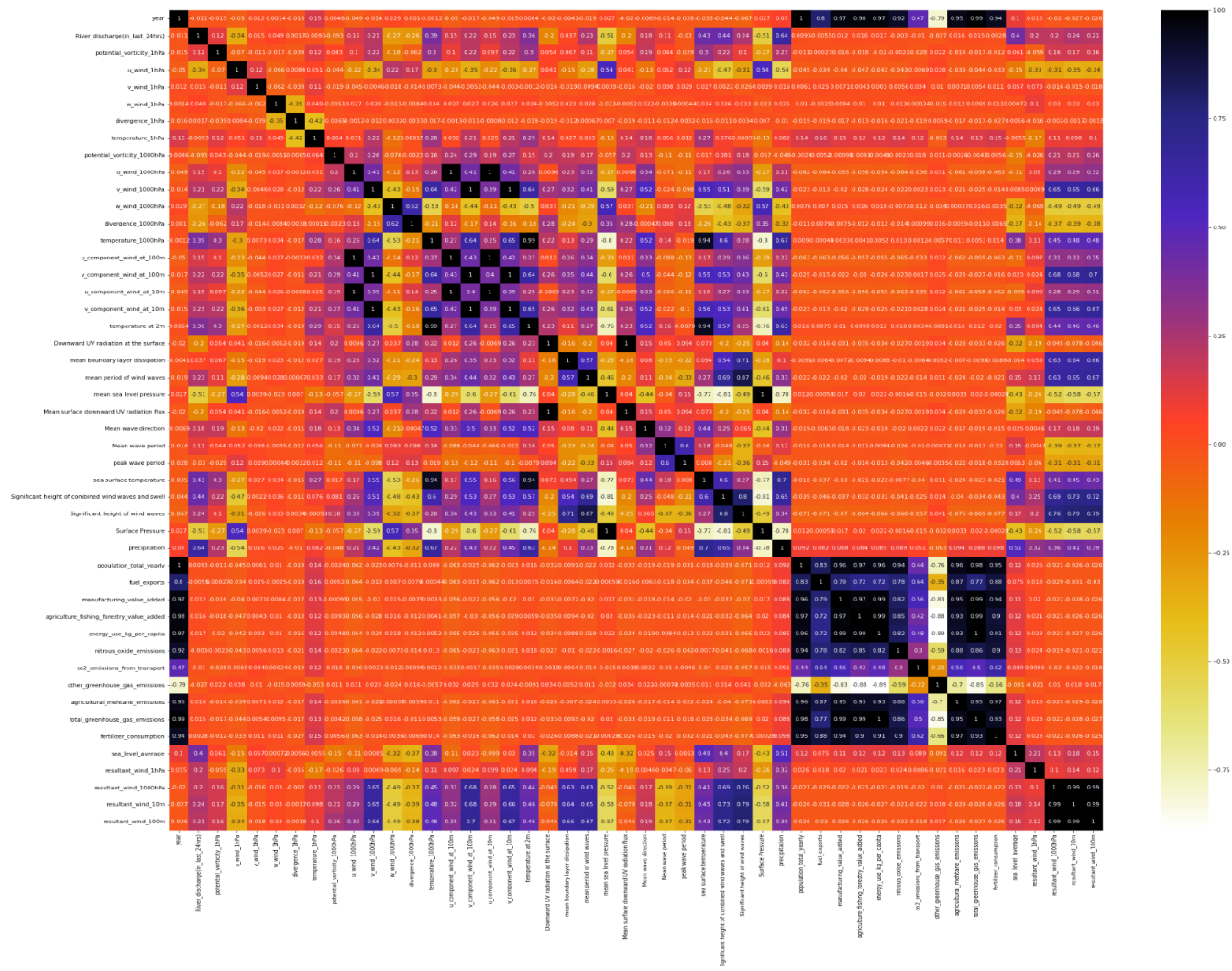

Fig 4.2.2.1 : Correlation heatmap with the raw data without dropping any feature

The above heatmap was generated in the process of Feature selection where higher contrast means higher correlation (positive or negative) For finding feature importance using ML, features having high correlation amongst each other were dropped in the process as they will only make our dataset bulkier.

But Note that while finding the causal factors, it is necessary to take all the variables into account.

The threshold was set to be 0.95 and it was found that the following variables were breaching that threshold and were thus removed.

1. temperature at 2m with temperature_1000hPa
2. Surface Pressure
3. Downward UV radiation at the surface
4. v_wind_1000hPa, v_component_wind_at_100m
5. u_wind_1000hPa, u_component_wind_at_100m
6. Population_total_yearly, manufacturing_value_added, agriculture_fishing_forestry_value added
7. Energy_use_kg_per_capita, nitrous_oxide_emissions, agricultural_mehtane_emissions

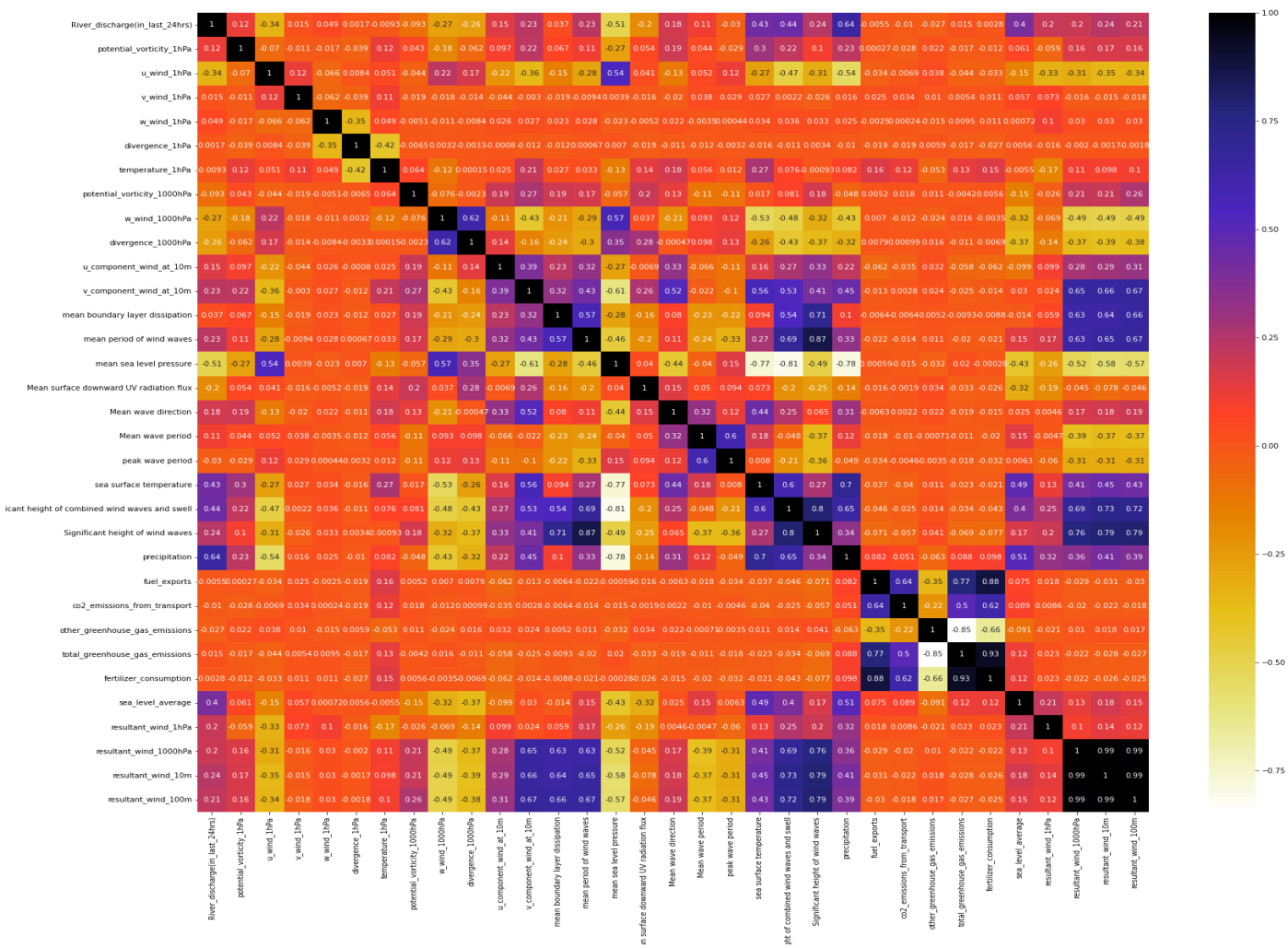


Fig 4.2.2.2 : Correlation heatmap generated after dropping intercorrelated variables

The above heatmap was generated after removing all the listed variables. The following indicators seemed fit and thus were included in our model as features-

1. **River_discharge(in_last_24hrs)** - ($m^3$ $s^{-1}$) It is the River discharge in the last 24 hours
2. **potential_vorticity_1hPa** - (K $m^2$ $kg^{-1}$ $s^{-1}$)Potential vorticity is a measure of the capacity for air to rotate in the atmosphere measured at 1hPa pressure.
3. **u_wind_1hPa** - This parameter is the eastward component of the wind at 1hPa pressure. It is the horizontal speed of air moving towards the east. A negative sign indicates air moving towards the west.
4. **v_wind_1hPa** - (m $s^{-1}$)This parameter is the northward component of the wind at 1hPa pressure level. It is the horizontal speed of air moving towards the north.
5. **w_wind_1hPa** - (Pa $s^{-1}$)This parameter is the speed of air motion in the upward or downward direction at 1hPa pressure level.
6. **divergence_1hPa** - ($s^{-1}$)This parameter is the horizontal divergence of velocity at 1hPa pressure level. It is the rate at which air is spreading out horizontally from a point, per square meter. This parameter is positive for air that is spreading out, or diverging, and negative for the opposite, for air that is concentrating.
7. **temperature_1hPa** - (K) This parameter is the temperature in the atmosphere at 1hPa pressure level. It has units of kelvin (K).
8. **potential_vorticity_1000hPa** - (K $m^2$ $kg^{-1}$ $s^{-1}$)Potential vorticity is a measure of the capacity for air to rotate in the atmosphere measured at 1000hPa pressure.
9. **w_wind_1000hPa** - (Pa $s^{-1}$)This parameter is the speed of air motion in the upward or downward direction at 1000hPa pressure level.
10. **divergence_1000hPa** - ($s^{-1}$)This parameter is the horizontal divergence of velocity at 1000hPa pressure level. It is the rate at which air is spreading out horizontally from a point, per square meter.
11. **u_component_wind_at_10m** - (m $s^{-1}$)This parameter is the eastward component of the wind, at a height of 10 meters above the surface of the Earth.
12. **v_component_wind_at_10m** - (m $s^{-1}$)This parameter is the northward component of the 10m wind.
13. **mean boundary layer dissipation** - ($Wm^{-2}$)This parameter is the mean rate of conversion of kinetic energy in the mean flow into heat, over the whole atmospheric column, per unit area, that is due to the effects of stress associated with turbulent eddies near the surface and turbulent orographic form drag.
14. **mean period of wind waves** - (s)This parameter is the average time it takes for two consecutive wave crests, on the surface of the ocean/sea generated by local winds, to pass through a fixed point.
15. **mean sea level pressure** - (Pa)This parameter is the pressure (force per unit area) of the atmosphere at the surface of the Earth, adjusted to the height of mean sea level. It is a measure of the weight that all the air in a column vertically above a point on the Earth's surface would have, if the point were located at mean sea level.
16. **Mean surface downward UV radiation flux** - (W $m^{-2}$)This parameter is the amount of ultraviolet (UV) radiation reaching the surface. It is the amount of radiation passing through a horizontal plane.
17. **Mean wave direction** - (degree true)This parameter is the mean direction of ocean/sea

surface waves.

18. **Mean wave period** - (s)This parameter is the average time it takes for two consecutive wave crests, on the surface of the ocean/sea, to pass through a fixed point.
19. **peak wave period** - (s)This parameter represents the period of the most energetic ocean waves generated by local winds and associated with swell. The wave period is the average time it takes for two consecutive wave crests, on the surface of the ocean/sea, to pass through a fixed point.
20. **sea surface temperature** - (K)This parameter is the temperature of sea water near the surface.
21. **Significant height of combined wind waves and swell** - (m)This parameter represents the average height of the highest third of surface ocean/sea waves generated by wind and swell. It represents the vertical distance between the wave crest and the wave trough.
22. **Significant height of wind waves** - (m)This parameter represents the average height of the highest third of surface ocean/sea waves generated by the local wind. It represents the vertical distance between the wave crest and the wave trough.
23. **precipitation** - (m)This parameter is the accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface.
24. **fuel_exports** - Percentage of fuel exports in total merchandise exports.
25. **co2_emissions_from_transport** - percentage CO2 emissions from transport in total fuel combustion
26. **other_greenhouse_gas_emissions** - (thousand metric tons of CO2 equivalent) Other greenhouse emissions from various sources such as HFC, PFC and SF6 .
27. **total_greenhouse_gas_emissions** - (thousand metric tons of CO2 equivalent) Total greenhouse emissions from all sources.
28. **fertilizer_consumption** - (kilograms per hectare of arable land)Amount of fertilizer consumed in kg per hectare of arable land.
29. **sea_level_average** - (m)Sea surface height above mean sea surface computed with respect to a 20-year mean reference period 1993 to 2012
30. **resultant_wind_1hPa** - (m $s^{-1}$) It is the total resultant wind speed at 1hPa Pressure level.
31. **resultant_wind_1000hPa** - (m $s^{-1}$) It is the total resultant wind speed at 1000hPa Pressure level.
32. **resultant_wind_10m** - (m $s^{-1}$) It is the total resultant wind speed at 10m above the surface of the earth.
33. **resultant_wind_100m** - (m $s^{-1}$) It is the total resultant wind speed at 10m above the surface of the earth.

It was made sure that there are no 'NaN' value present in the above dataset by the following method - scaled_df.isna().sum()

# 5
# METHODOLOGY

## 5.1 FOR PREDICTING SEA LEVEL

For predicting the sea level, ARIMA model was used which stands for Auto Regressive Integrated Moving Average. But to use ARIMA, we have to make sure that the data is stationary. There are many methods to check if the data is stationary, like Durbin-Watson test, Augmented Dickey Fuller test etc. When forecasting or predicting the future, most time series models assume that each point is independent of one another. The best indication of this is when the dataset of past instances is stationary. For data to be stationary, the statistical properties of a system do not change over time. This does not mean that the values for each data point have to be the same, but the overall behavior of the data should remain constant. Therefore, checking for stationarity using the two test which are as follows:

### 5.1.1 CHECKING FOR STATIONARITY

#### 5.1.1.1 ROLLING STATISTICS

In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also called a moving mean (MM) or rolling mean and is a type of finite impulse response filter. Variations include: simple, cumulative, or weighted forms. This method gives a visual representation of the data to define its stationarity. A moving variance or moving average graph is plotted and then it is observed whether it varies with time or not. In this method, a moving window of time is taken (taken as 365) and then the mean of that time period is calculated as the current value.

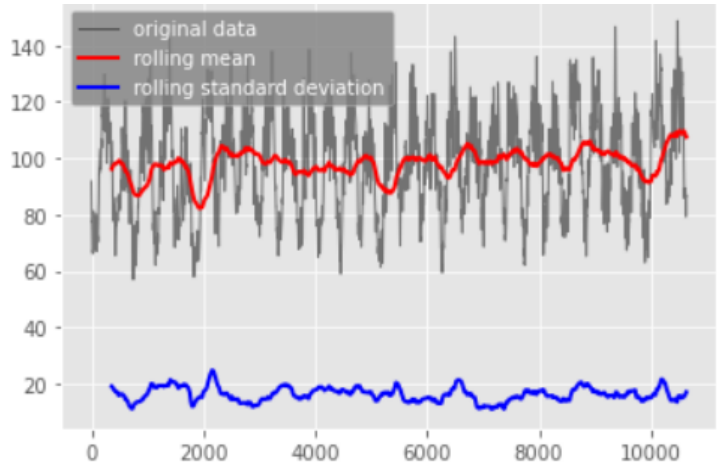


Fig 5.1.1.1.1 : Rolling mean and rolling std deviation

### 5.1.1.2 AUGMENTED DICKEY FULLER TEST

The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend. There are a number of unit root tests and the Augmented Dickey-Fuller may be one of the more widely used. It uses an autoregressive model and optimizes an information criterion across multiple different lag values.
The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.

- Null Hypothesis (H0) : If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
- Alternate Hypothesis (H1) : The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have a time-dependent structure.

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise a p-value above the threshold suggests we fail to reject the null hypothesis (non-stationary).

- p-value > 0.05 : Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- p-value <= 0.05 : Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

```
#we see that rolling mean and rolling statistics are
#Lets perform Dickey fuller test as well
from statsmodels.tsa.stattools import adfuller

dftest = adfuller(sla1, autolag = "AIC")
dfoutput = {}
dfoutput['Test statistics'] = dftest[0]
dfoutput['p-value'] = dftest[1]
dfoutput['no. of lags used'] = dftest[2]
dfoutput['number of observations used'] = dftest[3]
for key,value in dftest[4].items():
    dfoutput['critical value (%s)'%key] = value

dfoutput
```

```
{'Test statistics': -7.547898705387225,
 'p-value': 3.2466563221740574e-11,
 'no. of lags used': 36,
 'number of observations used': 10595,
 'critical value (1%)': -3.430967355831982,
 'critical value (5%)': -2.8618128362415436,
 'critical value (10%)': -2.5669152255899066}
```

Fig 5.1.1.2.1 : Test statistics for Dickey Fuller Test

We can see that the p-value is less than 0.05 and also the critical value is more than test statistics. Thus, we can conclude that Data is stationary and we can proceed.

### 5.1.2 ARIMA MODEL

The ARIMA model is an acronym for Auto-Regressive Integrated Moving Average, essentially creates a linear equation which describes and forecasts for the time series data. This equation is generated through three separate parts which can be described as:

- AR — auto-regression: equation terms created based on past data points
- I — integration or differencing: accounting for overall “trend” in the data
- MA — moving average: equation terms of error or noise based on past data points

Together, these three parts make up the AR-I-MA model.

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

Non-seasonal ARIMA models are generally denoted ARIMA(p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. Seasonal ARIMA models are usually denoted ARIMA(p,d,q)$(P,D,Q)_m$, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

When two out of the three terms are zeros, the model may be referred to based on the non-zero parameter, dropping ‘AR’, ‘I’ or ‘MA’ from the acronym describing the model. For example, ARIMA ( 1 , 0 , 0 ) is AR(1), ARIMA ( 0 , 1 , 0 ) is I(1), and ARIMA ( 0 , 0 , 1 ) is MA(1).

Given time series data $X_t$ where t is an integer index and the $X_t$ are real numbers, an ARIMA (p′,q) model is given by

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

Or equivalently by

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

where L is the lag operator, the $\alpha_i$ are the parameters of the autoregressive part of the model, the $\theta_i$ are the parameters of the moving average part and the $\varepsilon_t$ are error terms. The error terms $\varepsilon_t$ are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

Assume now that the polynomial $\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)$ has a unit root (a factor ( 1 − L )) of multiplicity d. Then it can be rewritten as:

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)=\left(1-\sum_{i=1}^{p'-d}\varphi_i L^i\right)(1-L)^d.$$

An ARIMA(p,d,q) process expresses this polynomial factorisation property with p=p'−d, and is given by:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

and thus can be thought of as a particular case of an ARMA(p+d,q) process having the autoregressive polynomial with d unit roots. (For this reason, no process that is accurately described by an ARIMA model with d > 0 is wide-sense stationary.)

The above can be generalized as follows.

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d X_t=\delta+\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t.$$

To get the values of p,q and d we need to draw ACF and PACF graphs respectively which stand for Autocorrelation function and partial autocorrelation function respectively.

**Autocorrelation Function (ACF)** - Autocorrelation is the correlation between a time series with a lagged version of itself. The ACF starts at a lag of 0, which is the correlation of the time series with itself and therefore results in a correlation of 1.

The ACF plot can provide answers to the following questions:

- Is the observed time series white noise/random?
- Is an observation related to an adjacent observation, an observation twice-removed, and so on?
- Can the observed time series be modeled with an MA model? If yes, what is the order?

**Partial Autocorrelation Function (PACF)** - The partial autocorrelation at lag k is the autocorrelation between X_t_t and X_(t-k) that is not accounted for by lags 1 through $k-1$.

The PACF plot can provide answers to the following question:

- Can the observed time series be modeled with an **AR model**? If yes, what is the order?

## 5.2 FOR ANALYSIS AND FINDING THE CAUSAL FACTORS BEHIND THE CHANGE

For finding the most important feature and the causal factor the methods which were used are Regression Analysis, XGBoost, Pearl and Granger Causality tests. There are many methods to check if the data is heteroscedastic in nature for example Breusch-Pagan Test, White's test etc. It is to be made sure that there is no multicollinearity present in the data as well for which OLS regression results were used as well as r_squared value from Regression analysis were also used.

### 5.2.1 REGRESSION ANALYSIS

RA is a strategy applied for modeling and examining numerical data. Regression analysis uses one or more independent variables to predict the target variable. Linear regression is a linear model, e.g. it assumes the target output variable(y) is linearly dependent on one or more input variable (x). More clearly, that the target variable can be predicted using a linear combination of independent variables. Linear regression uses a least squares function to model the observational data. The data model, which represents simple linear regression, can
be written as

$$Y = a1X1 + a2X2 + ... + anXn + e$$

where Y is the dependent variable (sea level average in this case), X1, X2, ...., Xn are the independent variables mentioned above, a1, a2, ....., an are the regression coefficients and e is the error term. The error term represents the "noise" or the random disturbances in the dependent variable which could not be explained and is treated as a random variable. The coefficients of independent input variables are determined to give the best fit model of the data. Generally, the least squares method is adopted to evaluate the best fitting model.

#### 5.2.1.1 FITTING THE REGRESSION MODEL

The initial input variable selection was based on correlation coefficient of the input variable and the sea level average. These correlation coefficients measure the degree of relation between the two variables. Variables having high correlation coefficient with the sla(sea level average) were selected and used for the modeling purpose. MS Excel and Python were used to carry out RA. To select the appropriate set of final input variables, backward elimination technique was carried out iteratively until we get the most significant model.

#### 5.2.1.2 AVOIDING MULTICOLLINEARITY

A linear regression model encounters the problem of Multicollinearity when some highly correlated independent variables are used in the regression model. Due to this, model interpretation becomes difficult and an overfitting problem arises. Highly correlated independent variables can cause significant change to one another with a small change in them and so the model results fluctuate significantly which may result in a highly unstable model and cannot be relied upon to get accurate predictions. A small change in the data or model will drastically

change model outputs which is undesirable. So, in order to have a reliable prediction of causal factors, some independent variables which were highly correlated were discarded to avoid multicollinearity.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      sea_level_average   R-squared:                       0.533
Model:                            OLS   Adj. R-squared:                  0.532
Method:                 Least Squares   F-statistic:                     364.0
Date:                Tue, 25 Apr 2023   Prob (F-statistic):               0.00
Time:                        03:42:01   Log-Likelihood:                -6086.3
No. Observations:               10220   AIC:                         1.224e+04
Df Residuals:                   10187   BIC:                         1.248e+04
Df Model:                          32
Covariance Type:            nonrobust
==========================================================================================================
                                                    coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                            -0.0669      0.009     -7.042      0.000      -0.085      -0.048
River_discharge(in_last_24hrs)                    0.0443      0.006      7.544      0.000       0.033       0.056
potential_vorticity_1hPa                         -0.0409      0.005     -7.593      0.000      -0.051      -0.030
u_wind_1hPa                                       0.0719      0.009      8.038      0.000       0.054       0.089
v_wind_1hPa                                       0.0169      0.005      3.419      0.001       0.007       0.027
w_wind_1hPa                                      -0.0142      0.005     -2.979      0.003      -0.024      -0.005
divergence_1hPa                                  -0.0249      0.005     -4.741      0.000      -0.035      -0.015
temperature_1hPa                                 -0.0773      0.007    -11.610      0.000      -0.090      -0.064
potential_vorticity_1000hPa                      -0.0049      0.001     -3.299      0.001      -0.008      -0.002
w_wind_1000hPa                                   -0.0205      0.009     -2.335      0.020      -0.038      -0.003
divergence_1000hPa                               -0.0697      0.007     -9.910      0.000      -0.083      -0.056
u_component_wind_at_10m                          -0.0563      0.006     -9.284      0.000      -0.068      -0.044
v_component_wind_at_10m                          -0.2272      0.014    -16.523      0.000      -0.254      -0.200
mean boundary layer dissipation                  -0.0569      0.004    -13.482      0.000      -0.065      -0.049
mean period of wind waves                        -0.0268      0.011     -2.532      0.011      -0.048      -0.006
mean sea level pressure                           0.0084      0.021      0.397      0.691      -0.033       0.050
Mean surface downward UV radiation flux          -0.1456      0.007    -19.928      0.000      -0.160      -0.131
Mean wave direction                              -0.0269      0.003     -9.499      0.000      -0.032      -0.021
Mean wave period                                  0.0758      0.010      7.456      0.000       0.056       0.096
peak wave period                                 -0.0131      0.008     -1.577      0.115      -0.029       0.003
sea surface temperature                           0.4377      0.014     31.843      0.000       0.411       0.465
Significant height of combined wind waves and swell  0.1115  0.019      5.962      0.000       0.075       0.148
Significant height of wind waves                 -0.0038      0.016     -0.245      0.807      -0.034       0.027
precipitation                                     0.1130      0.016      7.006      0.000       0.081       0.145
fuel_exports                                     -0.4348      0.033    -13.332      0.000      -0.499      -0.371
co2_emissions_from_transport                      0.0668      0.008      7.912      0.000       0.050       0.083
other_greenhouse_gas_emissions                    0.1524      0.013     11.793      0.000       0.127       0.178
total_greenhouse_gas_emissions                    0.6542      0.056     11.716      0.000       0.545       0.764
fertilizer_consumption                            0.0997      0.032      3.073      0.002       0.036       0.163
resultant_wind_1hPa                               0.0038      0.007      0.538      0.590      -0.010       0.017
resultant_wind_1000hPa                           -1.0359      0.084    -12.270      0.000      -1.201      -0.870
resultant_wind_10m                                0.8554      0.093      9.150      0.000       0.672       1.039
resultant_wind_100m                               0.3067      0.077      3.979      0.000       0.156       0.458
```

Fig 5.2.1.2.1 : OLS Regression Results

The *coef* is the coefficient ‘a’ of the regression model. The Adj. R-Squared a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model. In the above case, 53% of the variance in sea level change can be explained by the given variables.

### 5.2.1.3 TEST FOR HOMOSCEDASTICITY

The assumption of homoscedasticity (meaning “same variance”) is integral to linear regression models. Homoscedasticity depicts a condition in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) do not change significantly with change in values of an independent variable. When the

magnitude of the error term differs across values of an independent variable, the model is said to violate the assumption of homoscedasticity and is said to be heteroscedastic in nature.
Breusch-Pagan test and White's test were carried out for checking for homoscedasticity. *'Statsmodels'* library in Python is used to carry out this test.

- Null Hypothesis (H0): Homoscedasticity is present (the residuals are distributed with equal variance)
- Alternative Hypothesis (HA): Heteroscedasticity is present (the residuals are not distributed with equal variance)

5.2.1.4 TEST FOR AUTO-CORRELATION

Autocorrelation occurs when a given time series value highly correlates with a lagged version of itself over successive time intervals. It's conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods. '*Statsmodels'* library in python is used to check for Auto-correlation.

### 5.2.2 XGBOOST

XGBoost stands for *'Extreme Gradient Boosting'*. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. Boosting is an ensemble learning technique to build a strong classifier from several weak classifiers in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only control for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective. *'Xgboost'* library in Python is used to carry this out.

Decision trees create a model that predicts the label by evaluating a tree of if-then-else true/false feature questions, and estimating the minimum number of questions needed to assess the probability of making a correct decision. Decision trees can be used for classification to predict a category, or regression to predict a continuous numeric value.
A Gradient Boosting Decision Trees (GBDT) is a decision tree ensemble learning algorithm similar to random forest, for classification and regression. Ensemble learning algorithms combine multiple machine learning algorithms to obtain a better model.

Both random forest and GBDT build a model consisting of multiple decision trees. The difference is in how the trees are built and combined. Random forest uses a technique called bagging to build full decision trees in parallel from random bootstrap samples of the data set. The final prediction is an average of all of the decision tree predictions.

The term *"gradient boosting"* comes from the idea of *"boosting"* or improving a single weak

model by combining it with a number of other weak models in order to generate a collectively strong model. Gradient boosting is an extension of boosting where the process of additively generating weak models is formalized as a gradient descent algorithm over an objective function. Gradient boosting sets targeted outcomes for the next model in an effort to minimize errors. Targeted outcomes for each case are based on the gradient of the error (hence the name gradient boosting) with respect to the prediction.

GBDTs iteratively train an ensemble of shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model. The final prediction is a weighted sum of all of the tree predictions. Random forest *"bagging"* minimizes the variance and overfitting, while GBDT *"boosting"* minimizes the bias and underfitting.

XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT. It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set.

#### 5.2.2.1 BUILDING THE MODEL

XGBoost takes many parameters such as '*n_estimators'* which is the number of boosting rounds to be performed. Similarly, the other parameters include '*min_child_weight'*, '*max_depth'* which is the maximum tree depth for base learners, '*learning_rate'*, '*gamma'*, '*colsample_bytree'* and so on. For the initial model the following parameters were given

```
xgb = XGBRegressor(n_estimators=100)
```

However, the model can be improved by changing the parameters to the optimal one. Hyperparameter tuning is a critical step in optimizing the performance of a machine learning model. XGBoost is no exception, and it has several hyperparameters that can be tuned to improve its predictive accuracy. Here's an overview of the hyperparameters in XGBoost and some common methods for tuning them.

1. Learning rate: Also known as the "shrinkage" parameter, this hyperparameter controls the step size at each iteration of the boosting process. A lower learning rate can help prevent overfitting, but it may also increase the number of iterations needed to converge. Common values for the learning rate are 0.1, 0.01, and 0.001.
2. Maximum depth: This hyperparameter controls the maximum depth of each decision tree in the ensemble. A deeper tree can capture more complex interactions in the data, but it can also increase the risk of overfitting. Common values for the maximum depth are between 3 and 10.
3. Minimum child weight: This hyperparameter controls the minimum sum of instance

weights needed to split a node. A higher value can prevent the model from creating nodes with too few instances, which can help reduce overfitting.

4. Gamma: This hyperparameter controls the minimum reduction in loss required to make a further partition on a leaf node of the tree. A higher value can help prevent overfitting by reducing the number of splits.
5. Subsample: This hyperparameter controls the fraction of instances to be randomly sampled for each tree. A lower value can help prevent overfitting by reducing the variance of the model, but it can also reduce the model's ability to capture complex patterns in the data.

There are several methods for tuning hyperparameters in XGBoost, including grid search, random search, and Bayesian optimization. Grid search involves specifying a range of values for each hyperparameter and trying every possible combination. Random search involves randomly sampling from the range of values for each hyperparameter. Bayesian optimization involves using a probabilistic model to predict the performance of each hyperparameter configuration and selecting the configuration with the highest expected improvement.

In practice, it's often best to use a combination of these methods, along with cross-validation, to ensure that the hyperparameters are tuned properly. Cross-validation involves dividing the data into multiple folds and training the model on each fold while evaluating its performance on the remaining folds. This process helps ensure that the hyperparameters are tuned to generalize well to new data.Overall, hyperparameter tuning is an important step in optimizing the performance of an XGBoost model, and it requires careful consideration of the trade-offs between bias and variance, as well as the available computational resources.

*GridSearchCV* and *RandomizedSearchCV* were used from the 'sklearn.model_selection library' to find the best parameters. The parameters were finally found out to be following:

'n_estimators' = 200,
'min_child_weight' = 1,
'max_depth' = 100,
'learning_rate' = 0.25,
'gamma' = 0.0,
'colsample_bytree' = 1.

By using the above parameters, the Mean Squared Error was reduced from 0.009455952898550874 to 1.540575623470922e-07 which is a significant reduction.

5.2.2.2 CALCULATING FEATURE IMPORTANCE

There are several types of importance in the XGBoost - it can be computed in several different ways. The default type is gain. The gain type shows the average gain across all splits where the feature was used. Information Gain is simply the expected reduction in entropy caused by partitioning the examples according to this attribute.

$IG(S, a) = H(S) – H(S | a)$

where, IG(S, a) is the information for the dataset S for the variable a for a random variable, H(S) is the entropy for the dataset before any change and H(S | a) is the conditional entropy for the dataset given the variable a.

The feature_importances_ attribute of the model after training returns the feature importances for the problem.

### 5.2.3 CAUSAL INFERENCE

Causal inference is the process of ascribing causal relationships to associations between variables. Statistical inference is the process of using statistical methods to characterize the association between variables. Causality is at the root of scientific explanation which is considered to be causal explanation. However, establishing causal relationships is extremely difficult in spite of substantial advancements made during the past decades. Statistical inference works like a black box and generates the best possible characterization of the relationships between variables. Statistical inference provides estimates of the associations between variables but of course, association does not imply causation, so there is little that statistical inference can provide to establish causation.

5.2.2.1 GRANGER CAUSALITY

Granger causality is a statistical concept of causality that is based on prediction. According to Granger causality, if a signal X1 "Granger-causes" (or *"G-causes"*) a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone. Its mathematical formulation is based on linear regression modeling of stochastic processes (Granger 1969). More complex extensions to nonlinear cases exist, however these extensions are often more difficult to apply in practice.

X is expressed as an autoregressive process (linear function of past values of itself)

$$X(t) \sim a1X(t-1) + a2X(t-2) + a3X(t-3) + .....$$

Again, the values of Y are used to estimate the value of X(t)

$$X(t) \sim a1X(t-1) + a2X(t-2) + a3X(t-3) + ..... + b1Y(t-1) + b2Y(t-2) + b3Y(t-3) + .....$$

If the estimation improves, Y Granger-causes X.
*"Statsmodels"* library in Python is used to carry out this test.

### 5.2.2.2 PEARL CAUSALITY

Pearl's work is based on 'Structural Causal Models', which is a triple M = (U, V, F). In this model U is the collection of the exogenous (background, or driving) unobserved variables, V is the collection of endogenous (determined in some way by variables from U and V) variables, and F is a collection of functions f1, f2, ..., for each Vi in V. The variable Vi is fully determined as Vi = fi(U, V \ Vi), that is the arguments to fi are some of the variables in U, and some of the variables in V, but not Vi itself. In order to turn this into a probabilistic model, U is augmented with a probability distribution. An example is given where U1 is a court order for a man's execution, V are the actions of a captain (V1) and two riflemen (V2,V3) in a firing squad as well as the living/dead state of the person to whom the court order pertains (V3). If the judge orders the man shot (U1 = 'execute'), then this causes the captain to issue the order to fire, which causes the riflemen to shoot the prisoner, and hence causing his death. If the court order is not given, the captain remains silent, the riflemen don't shoot, and the prisoner is left alive.
To define the causal effect, consider two worlds: 1. World 1 (Real World): Where the action A was taken and Y observed 2. World 2 (Counterfactual World): Where the action A was not taken (but everything else is the same) Causal effect is the difference between Y values attained in the real world versus the counterfactual world.
In other words, A causes Y iff changing A leads to a change in Y, keeping everything else constant. Changing A while keeping everything else constant is called an intervention, and represented by a special notation do(A). Formally, causal effect is the magnitude by which Y is changed by a unit interventional change in A:

$$CE = E(Y|do(A=1)) - E(Y|do(A=0))$$

A CE value of >=0.05 is taken as the threshold for A to have a causal effect on Y for our analysis,
"*Dowhy"* library in Python is used to carry out this test.

### 5.2.2.3 STRUCTURAL CAUSAL MODEL

As defined above, the formation of the Structural Causal Model for Pearl Causality requires extensive domain knowledge. As it is beyond the scope of this project, the confounding variables

for the model are estimated using the correlation between the treatment and outcome variables. Variables with high correlation with both the treatment and outcome variables are taken as the common cause variables and those with high correlation only with the treatment variable are taken as the instrument variables.

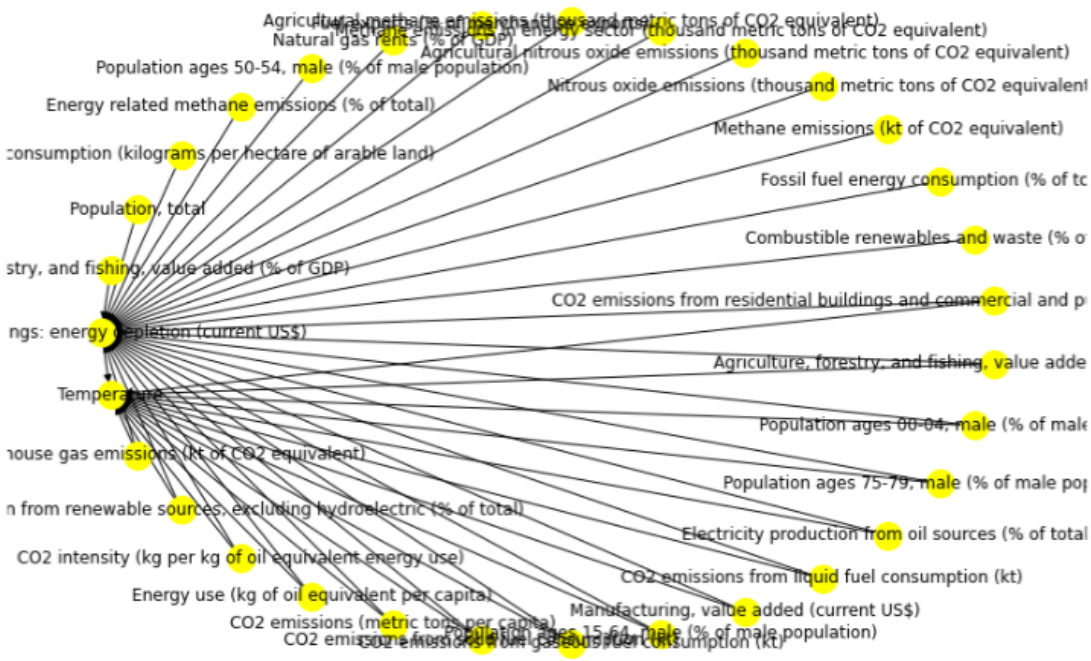


Fig 5.2.2.3.1 : Structural Causal Model

# 6
# RESULTS AND SUMMARY

## 6.1 PREDICTION RESULTS AND SUMMARY

We found out that the data is stationary and is thus suitable for the ARIMA model. Therefore, we proceed with the model. I have used “tsa.seasonal” module of statsmodel to decompose data into 3 parts which are

- Trend - It is the overall direction of the data
- Seasonality - It is the periodic component of the data
- Residual - It is the random fluctuations/noise which are caused by rare events and cannot be modeled.

A generated figure depicting the same is shown below.

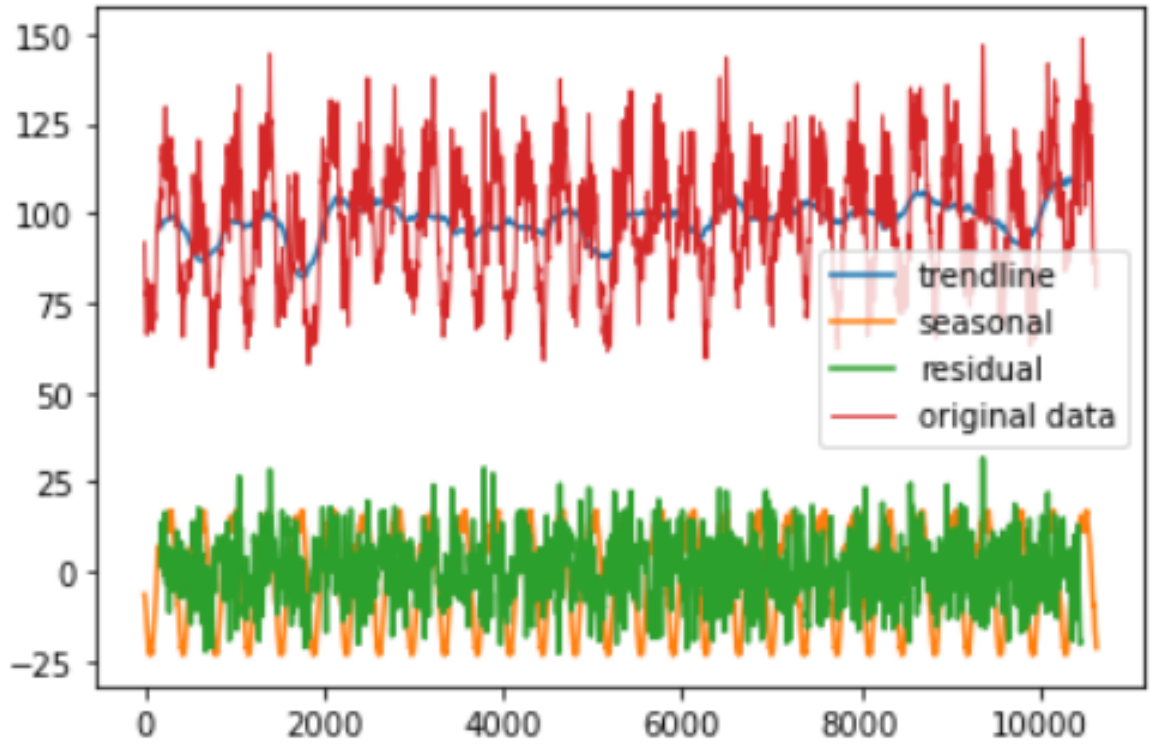


Fig 6.1.1 : Original data along with its decomposed components

Lets see these 3 components one by one

1. Trend Line

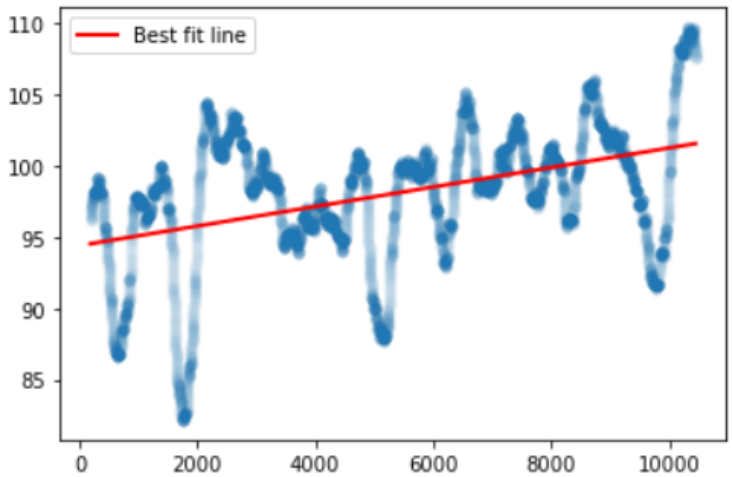


Fig 6.1.2 : Trendline and the best fit line of the same

2. Seasonality

Seasonality graph for first 1700 days is plotted as shown

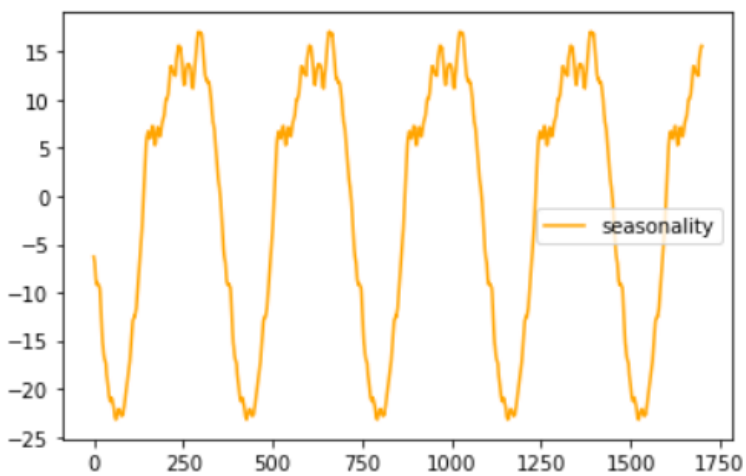


Fig 6.1.3 : Seasonality graph for the first 1700 days

We can observe from the above figures that the trend is upwards and our data has a seasonal component to it i.e. sea level is maximum at around 250th day which is in the month of August to September and it dips in the month of January and February.

Having seen this trend in our data, we can now proceed to generating ACF and PACF plots

which will give us our p and q values respectively using which we can generate future predictions.

The Autocorrelation and Partial Autocorrelation graph is generated by the 'stattools' module of the 'statsmodels.tsa' package. Figures of both is shown below

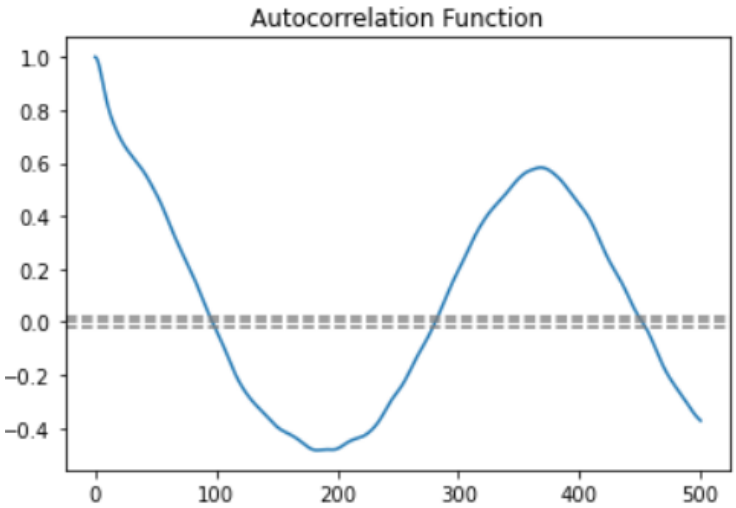


Fig 6.1.4 : ACF Plot

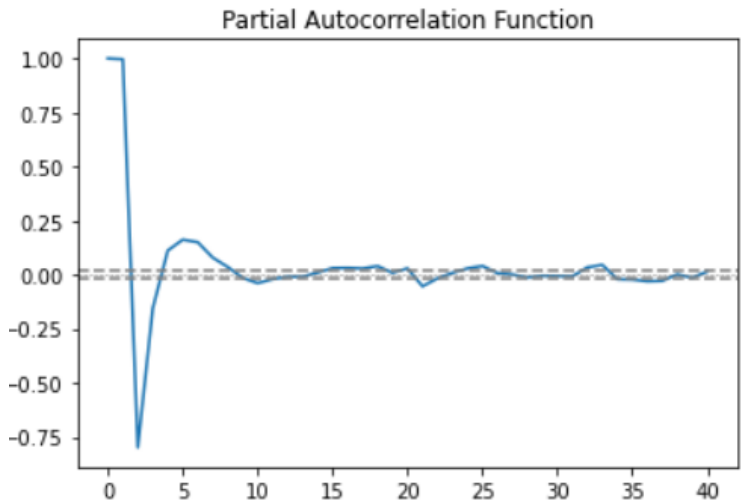


Fig 6.1.5 : PACF Plot

We can observe from the ACF graph that the values are highly positively correlated with the lags between 320 to 380 i.e. the data is having correlation with the data that was around 350 days prior to that data point.

Now we have our p and q values which are 100 and 2 respectively. When fed into the ARIMA model, this should give us the desired prediction result but due to memory(computer) constraints we cannot model such a big value. Therefore, I have used monthly average data instead and have generated the prediction based on the monthly average data.

Generating a fitted plot alongside the original data plot.

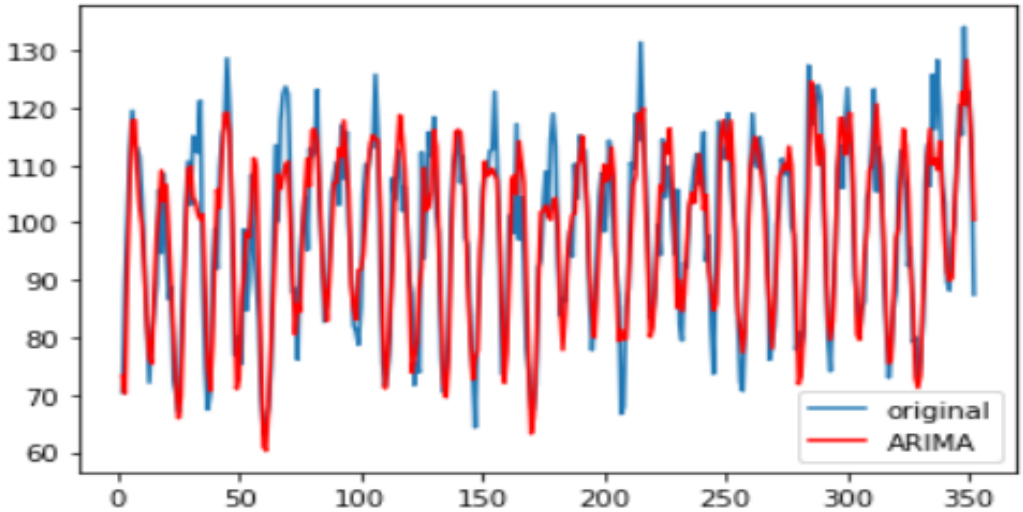


Fig 6.1.6 : Original along with the ARIMA fitted values

Now we use the '.predict' method to predict the values for the next 2 years. The original along with the predicted values graph is shown below

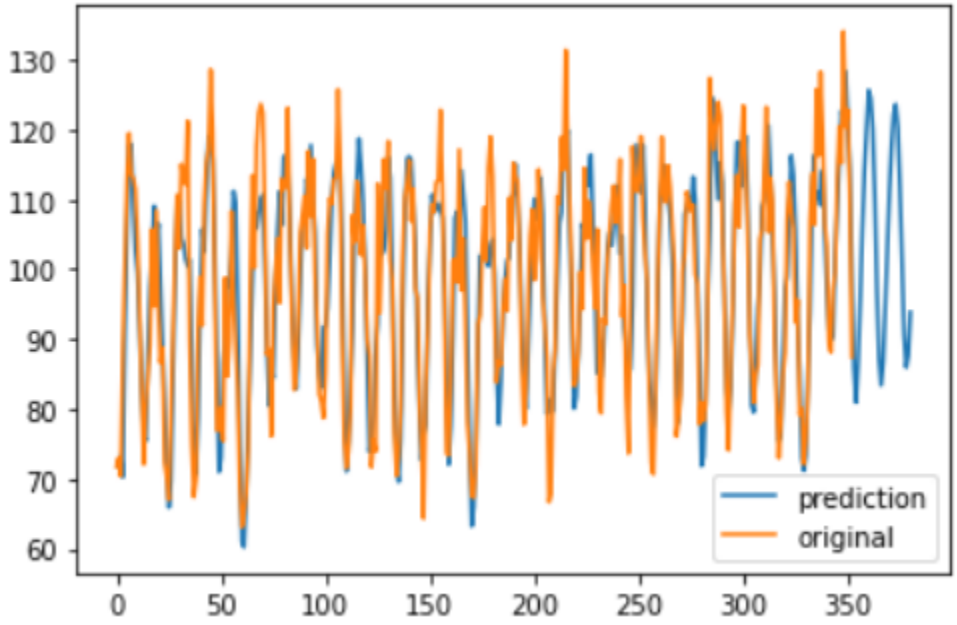


Fig 6.1.7 : Predicted and original sea level (month wise) from the year 1993 onwards

Long term prediction is generated using SARIMA function which stands for Seasonal ARIMA, it is able to capture Seasonal components with more precision. Long term prediction graph is generated with the level of confidence as shown below

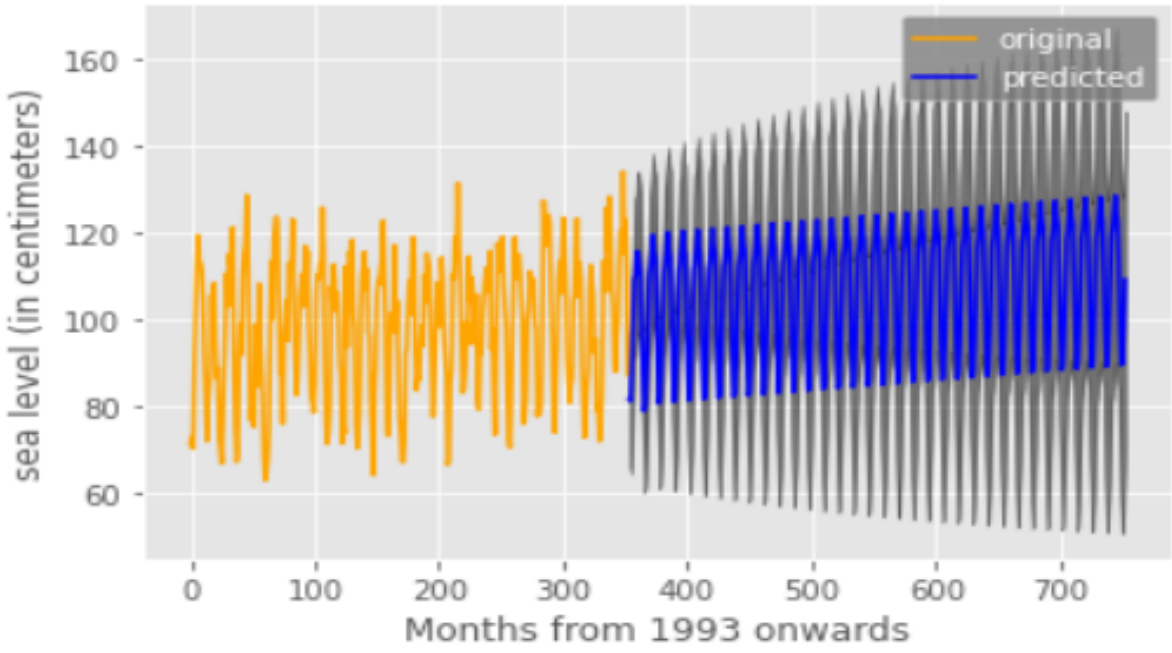


Fig 6.1.8 : Original and predicted sea level along with the 95% confidence level from the year 1993 onwards

## 6.2 FINDING THE CAUSAL FACTORS BEHIND THE RISE

### 6.2.1 REGRESSION ANALYSIS RESULTS

The results obtained after the regression analysis are as follows:

| | coef | std err | t | P>\|t\| |
|---|---|---|---|---|
| const | -0.0669 | 0.009 | -7.042 | 0.000 |
| River_discharge(in_last_24hrs) | 0.0443 | 0.006 | 7.544 | 0.000 |
| potential_vorticity_1hPa | -0.0409 | 0.005 | -7.593 | 0.000 |
| u_wind_1hPa | 0.0719 | 0.009 | 8.038 | 0.000 |
| v_wind_1hPa | 0.0169 | 0.005 | 3.419 | 0.001 |
| w_wind_1hPa | -0.0142 | 0.005 | -2.979 | 0.003 |
| divergence_1hPa | -0.0249 | 0.005 | -4.741 | 0.000 |
| temperature_1hPa | -0.0773 | 0.007 | -11.610 | 0.000 |
| potential_vorticity_1000hPa | -0.0049 | 0.001 | -3.299 | 0.001 |
| w_wind_1000hPa | -0.0205 | 0.009 | -2.335 | 0.020 |
| divergence_1000hPa | -0.0697 | 0.007 | -9.910 | 0.000 |
| u_component_wind_at_10m | -0.0563 | 0.006 | -9.284 | 0.000 |
| v_component_wind_at_10m | -0.2272 | 0.014 | -16.523 | 0.000 |
| mean boundary layer dissipation | -0.0569 | 0.004 | -13.482 | 0.000 |
| mean period of wind waves | -0.0268 | 0.011 | -2.532 | 0.011 |
| mean sea level pressure | 0.0084 | 0.021 | 0.397 | 0.691 |
| Mean surface downward UV radiation flux | -0.1456 | 0.007 | -19.928 | 0.000 |
| Mean wave direction | -0.0269 | 0.003 | -9.499 | 0.000 |
| Mean wave period | 0.0758 | 0.010 | 7.456 | 0.000 |
| peak wave period | -0.0131 | 0.008 | -1.577 | 0.115 |
| sea surface temperature | 0.4377 | 0.014 | 31.843 | 0.000 |
| Significant height of combined wind waves and swell | 0.1115 | 0.019 | 5.962 | 0.000 |
| Significant height of wind waves | -0.0038 | 0.016 | -0.245 | 0.807 |
| precipitation | 0.1130 | 0.016 | 7.006 | 0.000 |
| fuel_exports | -0.4348 | 0.033 | -13.332 | 0.000 |
| co2_emissions_from_transport | 0.0668 | 0.008 | 7.912 | 0.000 |
| other_greenhouse_gas_emissions | 0.1524 | 0.013 | 11.793 | 0.000 |
| total_greenhouse_gas_emissions | 0.6542 | 0.056 | 11.716 | 0.000 |
| fertilizer_consumption | 0.0997 | 0.032 | 3.073 | 0.002 |
| resultant_wind_1hPa | 0.0038 | 0.007 | 0.538 | 0.590 |
| resultant_wind_1000hPa | -1.0359 | 0.084 | -12.270 | 0.000 |
| resultant_wind_10m | 0.8554 | 0.093 | 9.150 | 0.000 |
| resultant_wind_100m | 0.3067 | 0.077 | 3.979 | 0.000 |

Fig 6.2.1.1 : Regression analysis result

The feature importance generated by Regression Analysis is shown above where the higher the 'coef' value the more is the correlation with the *'sea_level_average'*.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      sea_level_average   R-squared:                       0.533
Model:                            OLS   Adj. R-squared:                  0.532
Method:                 Least Squares   F-statistic:                     364.0
Date:                Tue, 25 Apr 2023   Prob (F-statistic):               0.00
Time:                        05:44:10   Log-Likelihood:                -6086.3
No. Observations:               10220   AIC:                         1.224e+04
Df Residuals:                   10187   BIC:                         1.248e+04
Df Model:                          32
Covariance Type:            nonrobust
```

Fig 6.2.1.2 : OLS Regression result continued

R-squared - It signifies the "*percentage variation in dependence that is explained by independent variables*". Here, 53.3% variation in 'sea level' is explained by all the 32 variables. This statistic has a drawback, it increases with the number of predictors(dependent variables) increasing. Therefore, Adj R-square is used which is adjusted for the number of variables in the regression. It increases only when an additional variable adds to the explanatory power to the regression.

Prob(F-statistics) - It tells the overall significance of the regression. This is to assess the significance level of all the variables together unlike the t-statistic that measures it for individual variables. The null hypothesis under this is "all the regression coefficients are equal to zero". Prob(F-statistics) depicts the probability of null hypothesis being true. As per the above results, probability is zero. This implies that overall the regressions are meaningful.

### 6.2.2 XGBOOST RESULTS

The initial XGBoost model with no tuned parameters produced the following results:

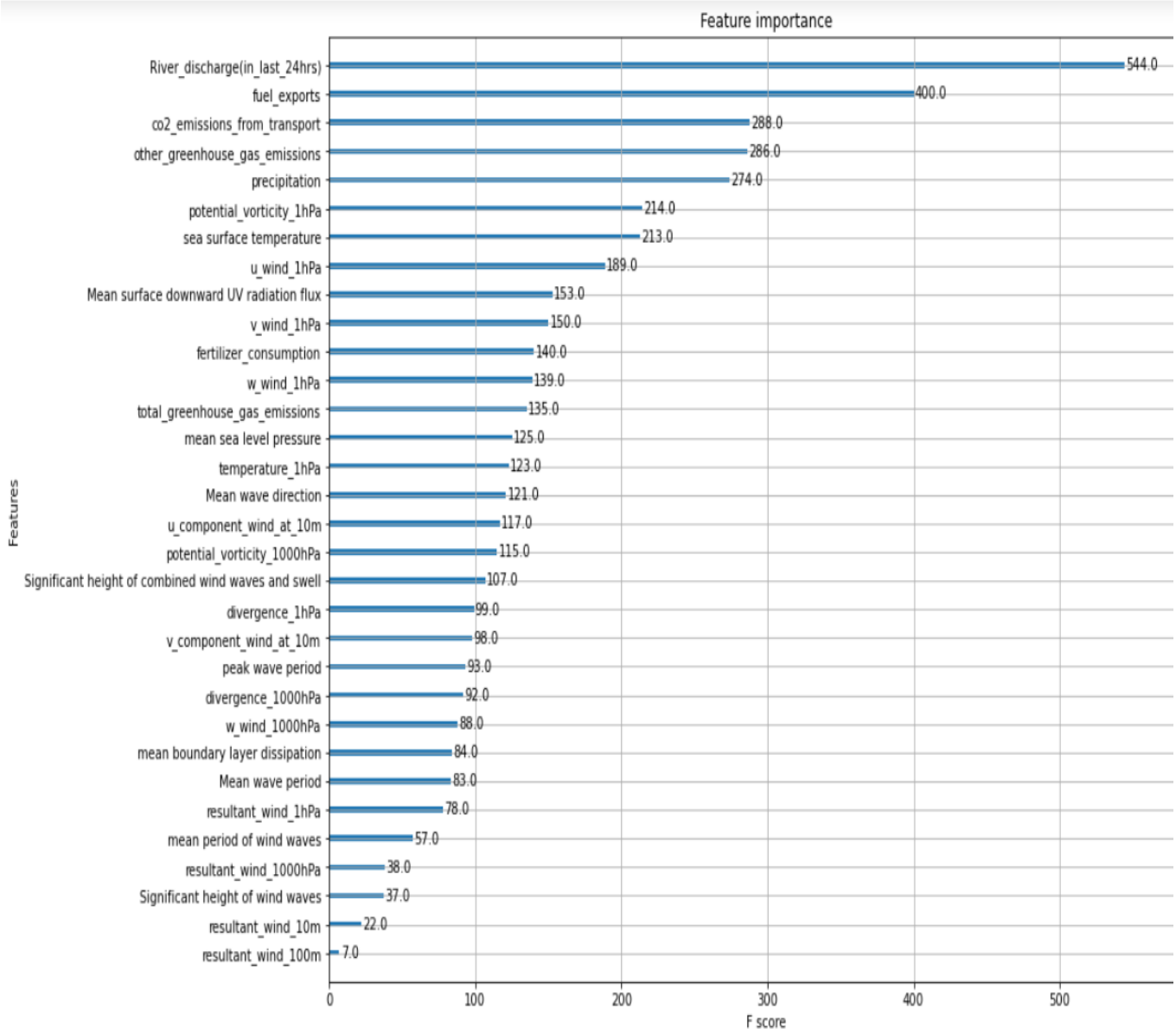


Fig 6.2.2.1: Feature importance graph by variables

We can see from the above chart that River discharge contributes most to the change in Sea level. Here the River discharge corresponds to the water discharge of the Ganges River near Haldia port. It is then followed by the fuel exports, CO2 emissions and other greenhouse gas emissions. The above observation can be thought of intuitively as well that the more water flowing into the sea, the more will be the sea level and the more the CO2 emissions, more will be the melting of ice and glaciers and thus more will be the sea level.

However, after Hyperparameter tuning, the result changes a bit which is depicted in the graph shown below.

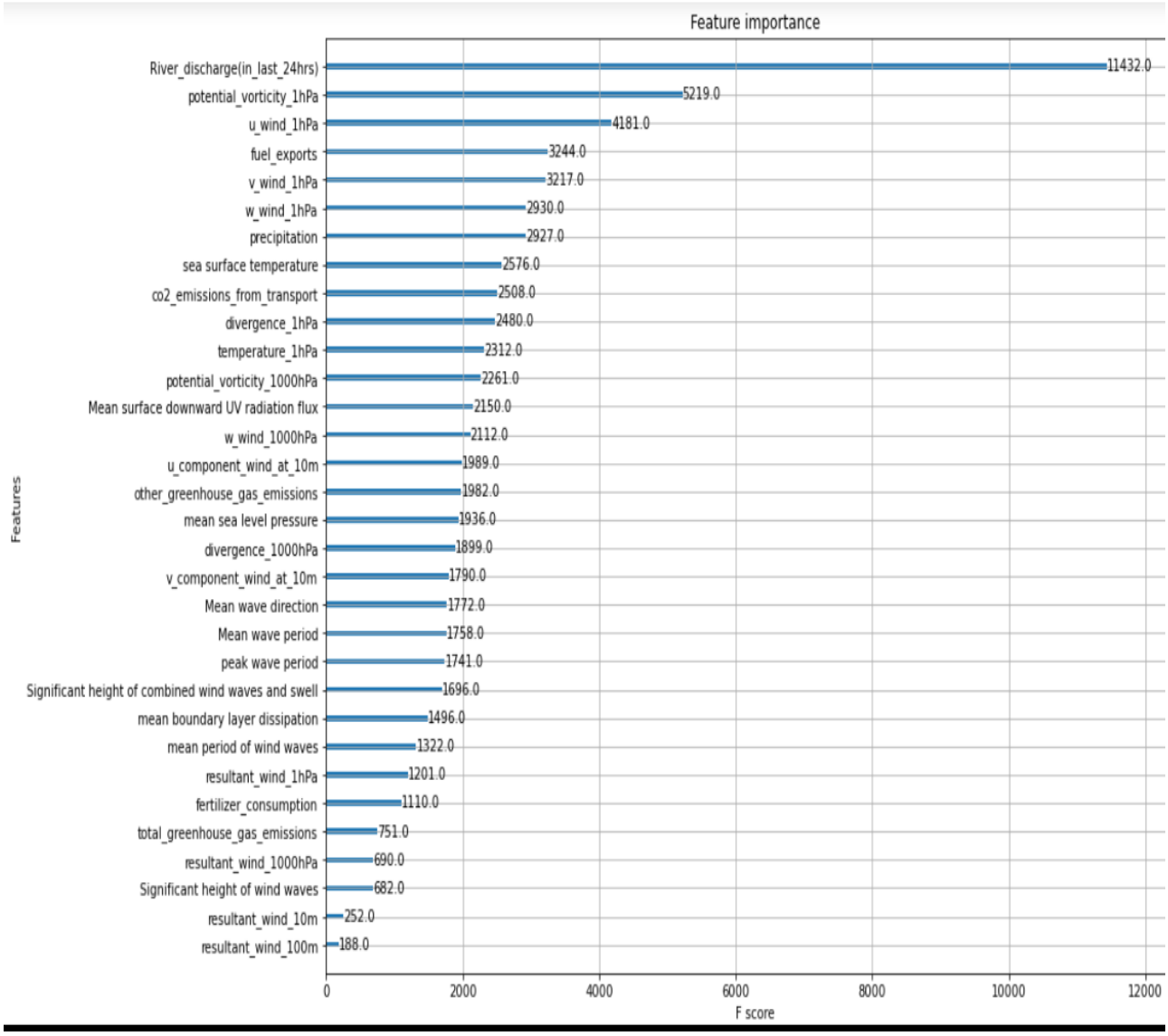


Fig 6.2.2.2 : Feature importance graph after hyper parameter tuning of the model

We can notice from the above graph that potential_vorticity at 1hPa along with u_wind, v_wind, w_wind at 1hPa moved up the order. Followed by fuel exports, sea surface temperature, precipitation and CO2 emissions which is in line with the current hypothesis that CO2 emissions are driving the sea level rise. But it is yet to be investigated whether it is the cause of the sea level rise. Also, the effects of u_wind, v_wind and w_wind at 1hPa pressure level on the sea level rise is to be investigated thoroughly.

If we plot River_discharge and sea_level average on top of each other, the figure looks something like this.

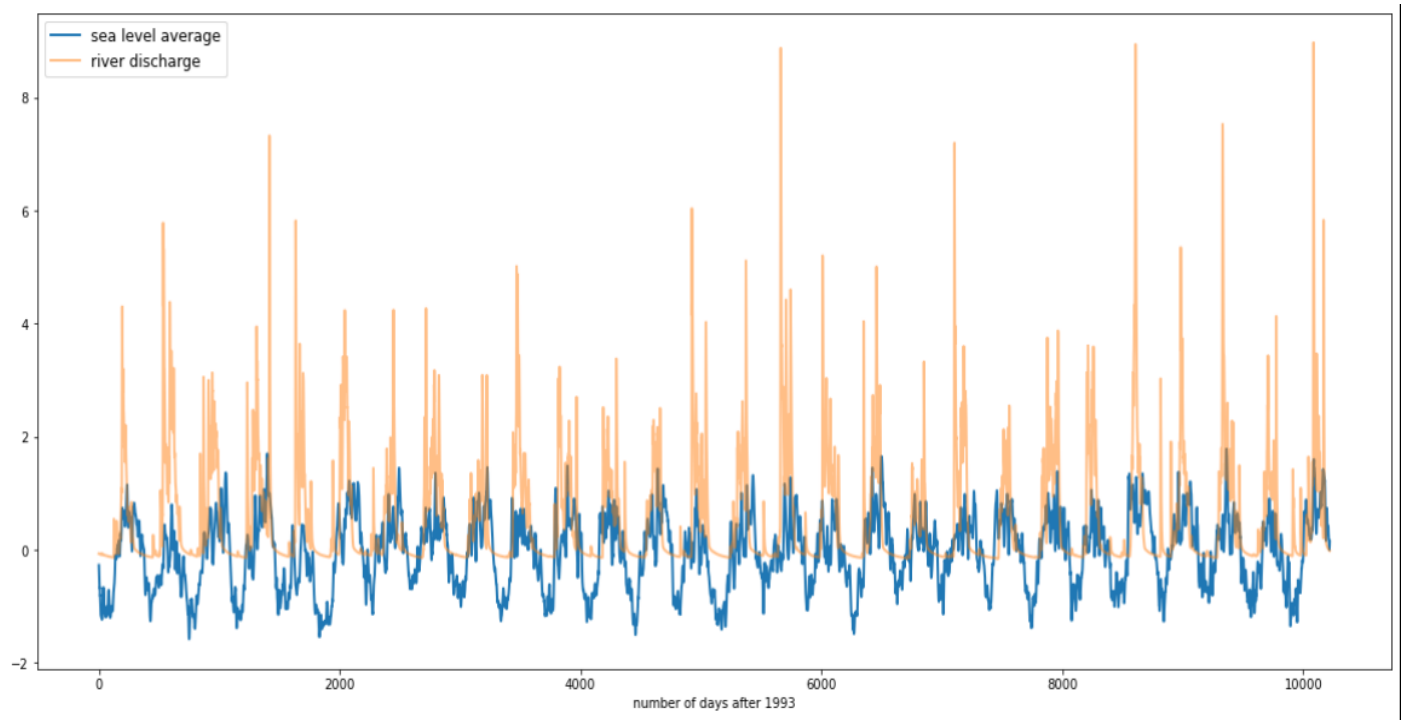


Fig 6.2.2.3 : Sea level average and the river discharge plotted together as a function of number of days

As we can see from the above figure, the peak of River discharge and sea level average coincides and thus it is considered to be the most important feature according to the XGBoost model.
If we plot vorticity at 1hPa with sea level average it is hard to notice any correlation with our naked eyes but after thorough analysis it can be seen that it is negatively correlated.

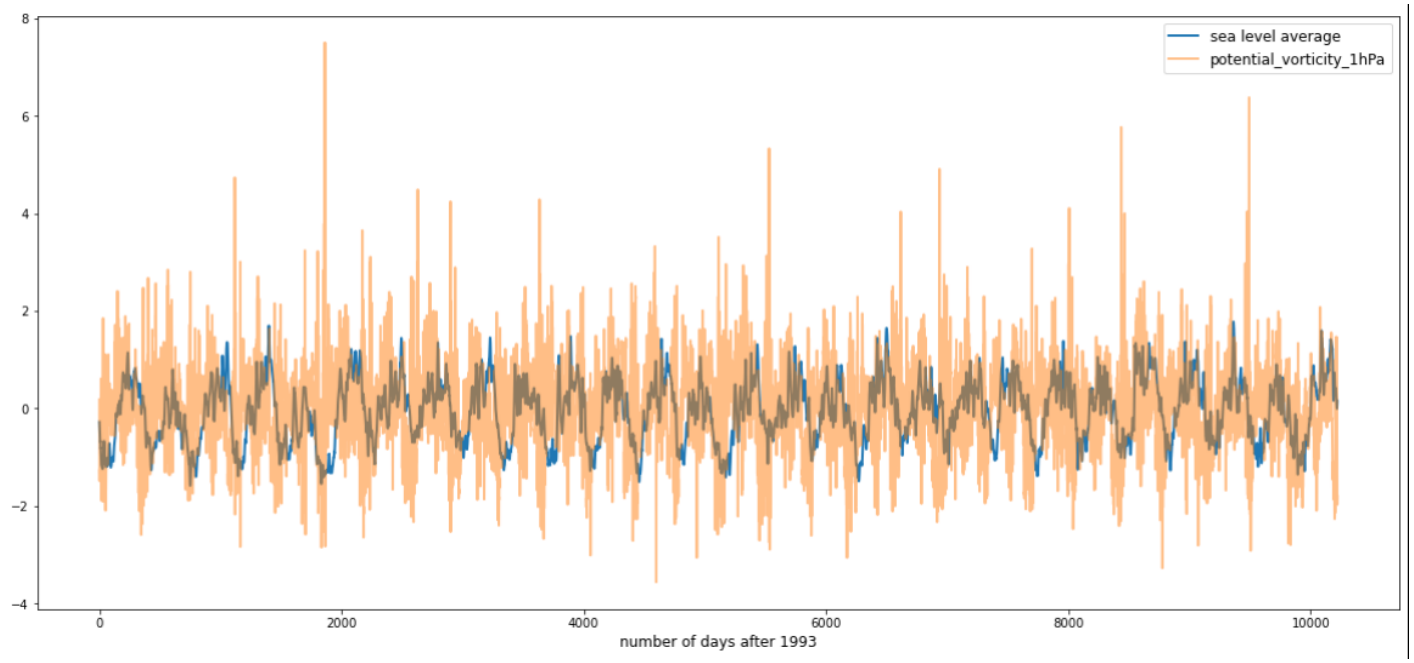


Fig 6.2.2.3 : Sea level average and the potenial_vorticity at 1hPa plotted together as a function of number of days

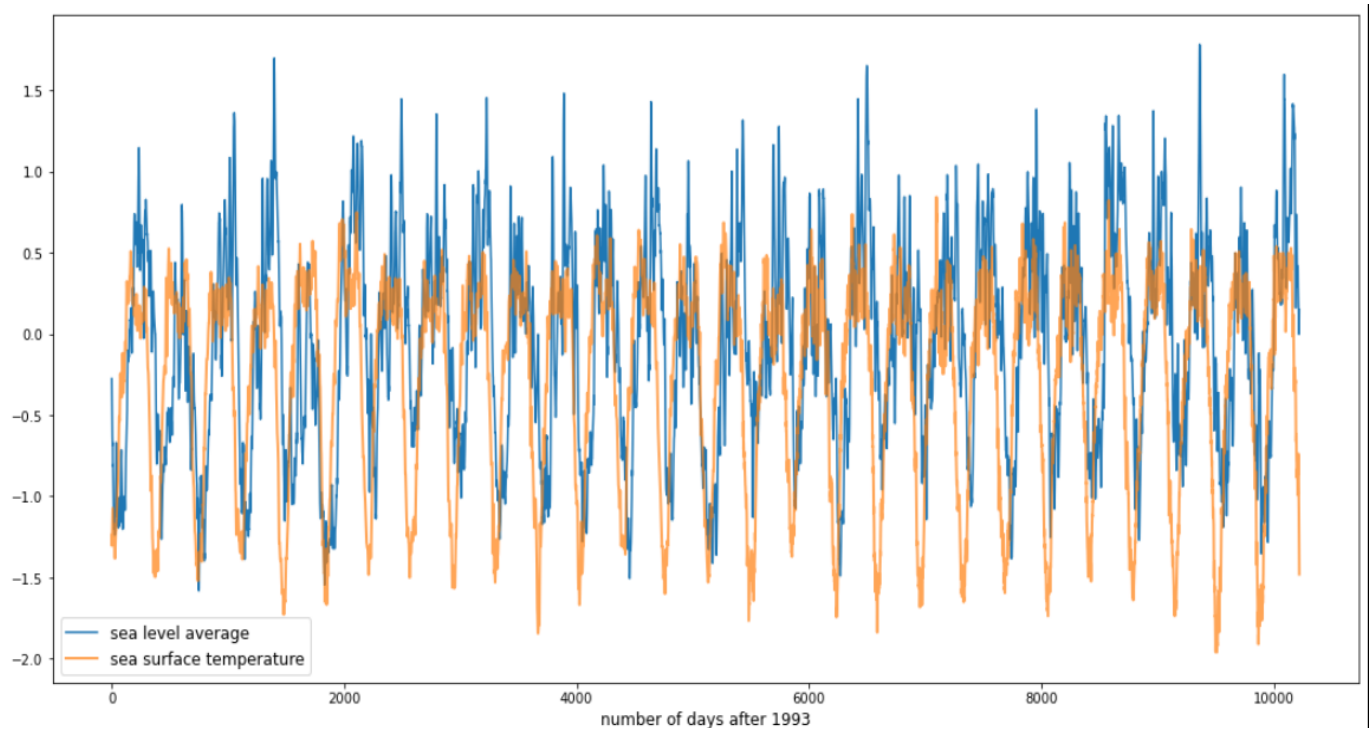


Fig 6.2.2.3 : Sea level average and the sea surface temperature plotted together as a function of number of days

From the above graph we can notice a positive correlation of the sea surface temperature with the sea level average. It is now time to investigate even further on what possibly could be the causal factors behind the sea level rise along with their lag. As for the Granger causality test, it produces the output along with their lag(years) by which the indicator is leading the 'Sea level average' which can give us an even more clear picture.

### 6.2.3 GRANGER CAUSALITY TEST RESULTS

We thus move on to the Granger Causality test results, which tell us the possible cause behind the sea level rise along with the time lag that one variable has over the other one. In a Granger causality test, time lag refers to the number of past observations of one time series that are included in the model to predict another time series. Specifically, the test examines whether past values of one time series provide useful information for predicting future values of another time series, after controlling for the effects of past values of the second time series.

The choice of time lag in a Granger causality test is an important consideration, as it can have a significant impact on the results. A lag that is too short may not capture the full relationship between the two time series, while a lag that is too long may include irrelevant or spurious relationships that can obscure the true causal relationship. For the test a lag of 36 months was taken.

The result produced are as follows:

| S. No | Variable | Lag (in months) |
|---|---|---|
| 1 | River_discharge(in_last_24hrs) | 1 |
| 2 | potential_vorticity_1hPa | 1 |
| 3 | potential_vorticity_1000hPa | 4 |
| 4 | u_wind_1000hPa | 2 to 5 |
| 5 | u_component_wind_at_100m | 2 to 5 |
| 6 | Mean boundary layer dissipation | 2 to 5 |
| 7 | Energy use per capita | 2 |
| 8 | CO2 emission from transport | 2 to 5 |
| 9 | Total greenhouse gas emissions | 2 to 5 |
| 10 | Resultant wind at 1hPa | 4 to 5 |

Therefore, the Granger causality test found 10 indicators which contributed most to the 'Sea level rise' along with the lag (in months) in between the effect and the cause. We can again see that River discharge tops the list with a maximum lag of 1 month. It is then followed by potential_vorticity at 1hPa which also has a lag of 1 month. We can also notice from the above table that low level wind does infact influence sea level which was not captured by the previous models since it has a lag of 2 to 5 months between effect and cause.
We can also conclude that Energy use per capita, CO2 emissions along with other greenhouse gas emissions influence the sea level with a minimum lag of 2 months and a maximum lag of 5 months.

## 6.3 CONCLUSION AND SUMMARY

We now have our sea level prediction for the next 33 months with the 95% confidence level(grey region) which tells us that the probability of the original value falling in this region is 95% and thus the name 95% confidence level. We can also observe that the confidence level is widening as the time passes by. It shows us that the predicted values may not be right but we can have an estimate of the series based on the data we have. We can also observe that the trend line captured by 'seasonal_decompose', though not very predictable, is certainly increasing as shown by the best fit line.

It was seen that the ARIMA model works beautifully on smaller dataset but it requires a heavy computational machine when the dataset is large and has high correlations with more than 10 days lag of itself i.e. p value is greater than 10. However, the results when assessed prove out to be quite robust for a single variable prediction. Use of SARIMA which stands for Seasonal ARIMA was also assessed using monthly average data and it outperformed ARIMA in catching seasonal trends and thus provided very reasonable long term predictions along with the 95% confidence level. Finally it is now firmly established that the sea level has risen in the last 3 decades and will continue to rise according to the prediction.

The effects of 44 variables were taken into account to assess the sea level change and it was found that the River Discharge from the Ganges was contributing most to the change in sea level rise at our target location since it is just a few kilometers south to the Haldia port. Therefore, it is just logical that a target location so close to the river basin is bound to affect the sea level. We also got to see that winds flowing at 1hPa pressure level along with potential vorticity has an effect on the sea level as well which was not expected. We also have seen how winds at different pressure levels affect the sea level.

$CO_2$ emissions along with other greenhouse gas emission effects were also seen with the help of the study. Unlike other natural factors, these are the one which are a direct cause of our lifestyle and it can be amended as well.

# 7
# FUTURE WORK

In this project, sea level analysis and projection were carried out for the coordinates 21.25°N,87.75°E from the period of 1993 to 2035 using the ARIMA model which needs only 1 variable. At times, it has some advantages but it is not reliable when we need accurate predictions. Therefore, prediction using multiple variables is recommended. Also, the study on the effect of anomalies such as hurricanes on sea level is yet to be seen which will be interesting nonetheless. Finally, causal factors behind sea level rise are to be found which will take much more effort and time and have to be carried out with precision.

As we know, volcanic eruptions in one part of the world can cause weather imbalance in some other parts of the world. Therefore, it is necessary to analyze the trends all over the world to come to any accurate conclusion about sea level rise. Thus, it is necessary to take a holistic view of the problem rather than viewing it as a country or a region specific problem. It would be quite a task to collect data on each country and mapping it in context of the whole world and it would require many more days to accomplish it but the result would be astonishing nonetheless.

Use of Neural networks, better causality tests after bringing domain specific knowledge and time series analysis is much welcomed in this type of study. For that, the need to collect relevant data across diverse fields along with the cleaning and taking care of missing data could also yield positive results.

It is much needed to therefore build a database for all the natural and human induced changes so that we can analyze, predict and mitigate the disaster that could be unfolding in the near future. For the above study, only 44 variables were taken into account to analyze the sea level. The number of variables could be easily taken to 200 or more which can give us a more clearer picture and can uncover a lot of insights that we might not see as of now. There is an urgent need to monitor the greenhouse gas emissions so as to get better results and we don’t have to rely on interpolated monthly data as we have done above.

# 8
# REFERENCES


1. Asteriou, Dimitros; Hall, Stephen G. (2011). "ARIMA Models and the Box–Jenkins Methodology". Applied Econometrics (Second ed.). Palgrave MacMillan. pp. 265–286.
2. Mills, Terence C. (1990). Time Series Techniques for Economists. Cambridge University Press.
3. Percival, Donald B., Walden, Andrew T. (1993). Spectral Analysis for Physical Applications. Cambridge University Press.
4. Leonie Monigatti,(Aug 3, 2022). Interpreting ACF and PACF Plots for Time series forecasting published on Towards Data Science, accessed on Nov 27, 2022.
5. Selva Prabhakaran, 2021. ARIMA Model - Complete guide to Time series forecasting, accessed on Nov 27, 2022.
6. Alan Bluis, (July 2020). Sea Level 101: part two: All sea level is local. NASA Climate Press. Accessed on Nov 27, 2022.
7. Miranda Ahul, (Aug 2021). What is an ARIMA Model ? published on Towards Data Science. Last Accessed on Nov 27, 2022.
8. M. Ali, R. Prasad, Y. Xiang, R.C. Deo. Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms, Renew. Sust. Ener. Rev., 132 (2020), Article 110003,
9. S. Asma, A. Sezer, O. Ozdemir. MLR and ANN models of significant wave height on the west coast of India. Computers & Geosciences, 49 (2012), pp. 231-237.
10. Intergovernmental Panel on Climate Change, in Climate Change 2013: The Physical Science Basis (eds Stocker, T. F. et al.) (Cambridge Univ. Press, Cambridge, 2013).
11. https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/
12. The official XGBoost documentation was referred to when trying to improve the model.