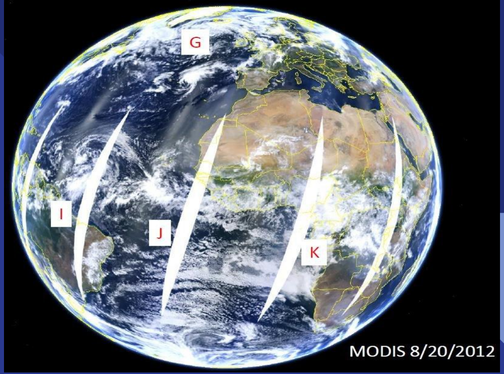



# Developing Spatial Simulator Engine For Air Pollutant Using Machine Learning

Presented by Debajyoti Maity

# Outline

- ❏ Introduction Of Modis Dataset
- ❏ Objective Of Our Study
- ❏ Data Source
- ❏ Data Description
- ❏ Data Extraction
- ❏ Data Prepossessing
- ❏ Exploratory Data Analysis
- ❏ Time Series Model building
- ❏ Conclusion

## Introduction Of Modis Dataset

- MODIS(Moderate Resolution Imaging Spectroradiometer) Aqua and Terra satellite data, which provide a wealth of high-resolution and multi-spectral information, this project seeks to create a sophisticated spatial engine capable of accurately detecting and mapping air pollutants on a global scale.

- Terra and Aqua are two Earth observing satellites launched as part of Earth Observing System(EOS) program. They are design to collect data about various aspects of Earth’s Atmosphere ,Oceans and Climate.

# Brief Discussion About Terra  and Aqua

- ## Terra (EOS AM-1):
  - Launch Date: December 18, 1999
  - Mission: Terra is equipped with a suite of instruments that observe Earth's atmosphere, land, oceans, and ecosystems. It's primarily focused on studying climate change, air quality, and environmental processes. Terra's observations help scientists monitor and understand the interactions between various components of the Earth system.
  - Key Instruments: Terra carries several instruments, including the Moderate Resolution Imaging Spectroradiometer (MODIS), the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), and the Clouds and the Earth's Radiant Energy System (CERES).
- ## Aqua (EOS PM-1):
  - Launch Date: May 4, 2002
  - Mission: Aqua is another component of the EOS program and is designed to study the Earth's water cycle, including water vapor, clouds, precipitation, and ice. Aqua's observations contribute to improving our understanding of the global energy and water.
  - Key Instrument :Aqua carries instruments like MODIS (which you mentioned), the Atmospheric Infrared Sounder (AIRS), and the Advanced Microwave Scanning Radiometer for EOS (AMSR-E), among others

## Graphical Representation

Aqua

The satellite (also called platform)

Terra

The instrument (also called sensor)

### Six Instruments

- MODIS
- CERES
- AIRS
- AMSU-A
- AMSR-E
- HSB

## Objective

- From the Satellite Image We have to Extract the Delhi Aod Dataset Then we training the Machine learning Model to predict $PM_{2.5}$ Of the Delhi.

# Why Chosen Delhi?

- Delhi was chosen as the project's focal point due to its notoriety for severe air pollution, which poses a significant health risk to its residents and demands innovative solutions. By predicting and addressing air quality issues in Delhi, we aim to improve public health and contribute to a cleaner and safe environment for the city's inhabitants.

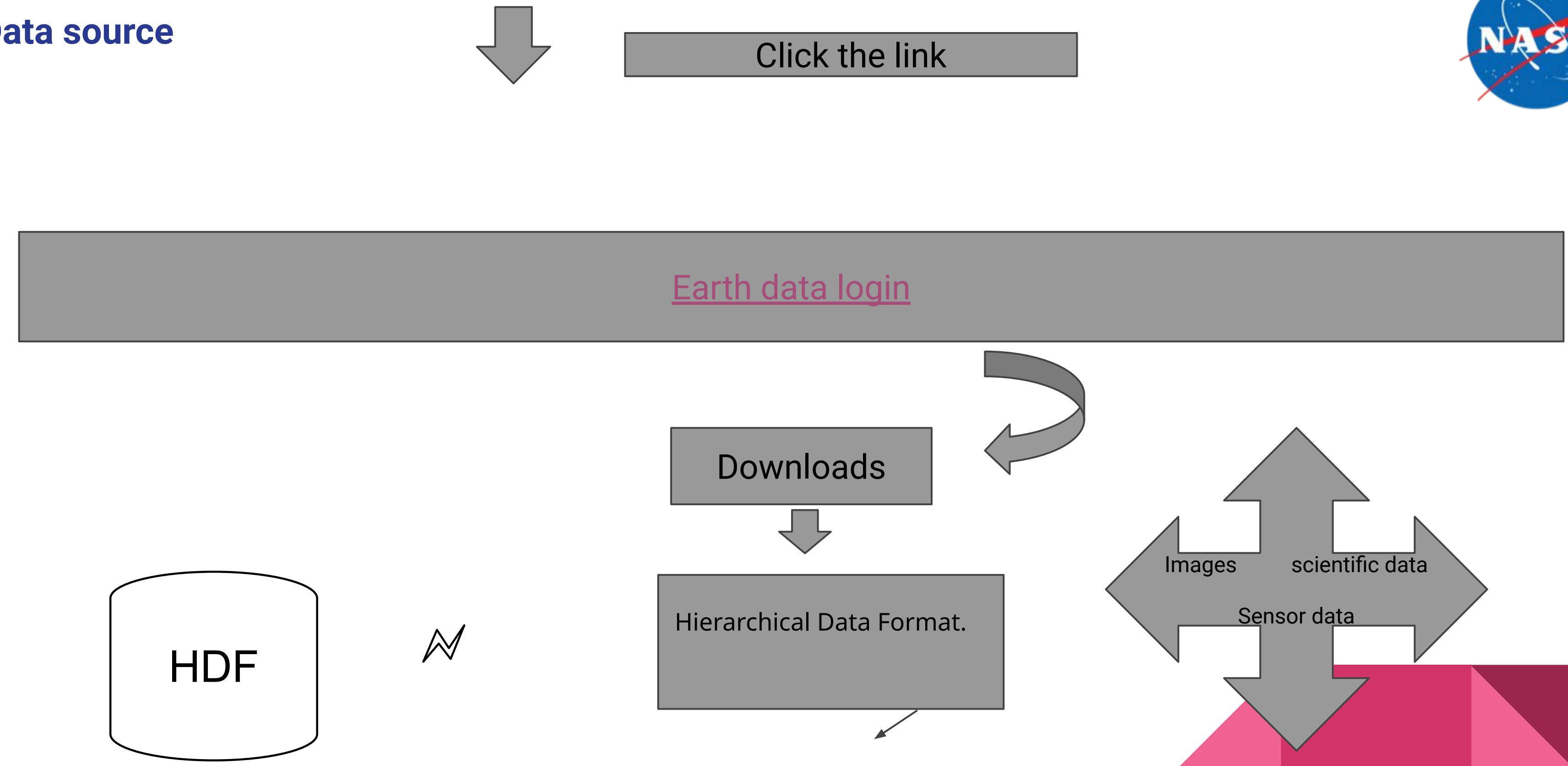
1)Improve public health
2)safe environment

# Description of AOD

- Aerosol Optical Depth (AOD) is a measure of the amount of aerosol particles present in the Earth's atmosphere and how much they affect the transmission of sunlight through the atmosphere. Aerosols are tiny solid or liquid particles suspended in the air, and they can include things like dust, smoke, pollen, pollutants, and sea salt.

- AOD is an important parameter for understanding atmospheric composition, air quality, climate, and visibility. It quantifies the degree to which aerosols scatter and absorb sunlight, affecting the amount of direct and diffuse solar radiation that reaches the Earth's surface. A high AOD indicates a higher concentration of aerosol particles and can lead to reduced visibility, altered energy balance, and potential cooling effects on the climate

## Data source


Click the link
Earth data login
Downloads
Hierarchical Data Format.
HDF
Images
scientific data
Sensor data

# HDF FORMAT

Grid 5km:

i)Data Fields
ii)Data attribute

i)Scattering angle
ii)Grid Angle
iii)cosSZA

Grid 1 Km

i)Data field
ii)Data attribute

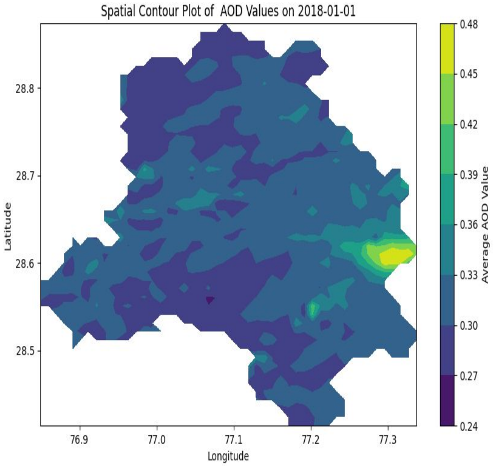
1)Optical_Depth_047
2)Optical_Depth_055
3)AOD_Uncertainty
4)Injection_Height
5)Column_WV
6)AngstromExp_470
7)AOD_QA

# Spatial Simulator Engine

features

Meteorological Data

1)Temperature
2)Precipitation
3)Humidity
4)Wind Speed and Direction
5)Pressure
6)Cloud Cover
7)Solar Radiation
8)Weather Phenomena
9)Air Quality




AOD heatmap

Satellite Image

Model

f

Prediction $PM_{2.5}$
Next time stamp

# Data Extraction

- Construct Coordinate  From hdf format
- Construct the boundary point of the Delhi Shapefile
- Extract the intermediate 1350 point from this shapefile
- Then extract the AOD value of this 1350 points by using Haversine formula
- Then this whole process is done in the year of 2018

# Delhi Shapefile Description

- ❖ Delhi Maximum Latitude : 28.883495792 $^0$N
- ❖ Delhi Minimum Latitude : 28.40425221 $^0$ N
- ❖ Delhi Maximum Longitude :77.3474 70 $^0$E
- ❖ Delhi Minimum Longitude : 76.838772 $^0$E

# Haversine Formula

The Haversine formula is a fundamental mathematical tool used to calculate the distance between two points on the Earth's surface given their latitude and longitude coordinates. It's particularly useful for determining the closest geographic point in a dataset to a user-specified location. Here's how the formula works:

- **Define Earth's Radius**: First, we define the radius of the Earth in meters as R (approximately 6,371,000 meters), as Earth is not a perfect sphere, and this value is an average radius.
- **Convert Latitude to Radians:** Latitude values need to be converted from degrees to radians because trigonometric functions work with radians. For both the user's location and the dataset points.
- **Convert the latitude from degrees to radians using np.radians(lat)**.(1)let consider latitude,longitude are the extract from hdf dataset (2) user_lat comes from delhi shapefile dataset (latitude) (3)user_lon is the longitude comes from delhi shapefile dataset
- **Calculate Differences in Latitude and Longitude**: Compute the differences between the user's latitude and the dataset's latitude, as well as the longitude differences between the user's location and the dataset points:
- 1)delta_lat = np.radians(latitude - user_lat) → This gives the radian of difference of the latitude and user's latitude

2)delta_lon = np.radians(longitude - user_lon) → This gives the radian of difference of the longitude and user's longitude

3)lat1 = np.radians(user's latitude) 4)lat2 = np.radians(latitude)

- **Intermediate Calculations (a):** Calculate intermediate values a using trigonometric functions and the differences in latitude and longitude:

a = (np.sin(delta_lat/2))**2 + (np.cos(lat1)) * (np.cos(lat2)) * (np.sin(delta_lon/2))**2

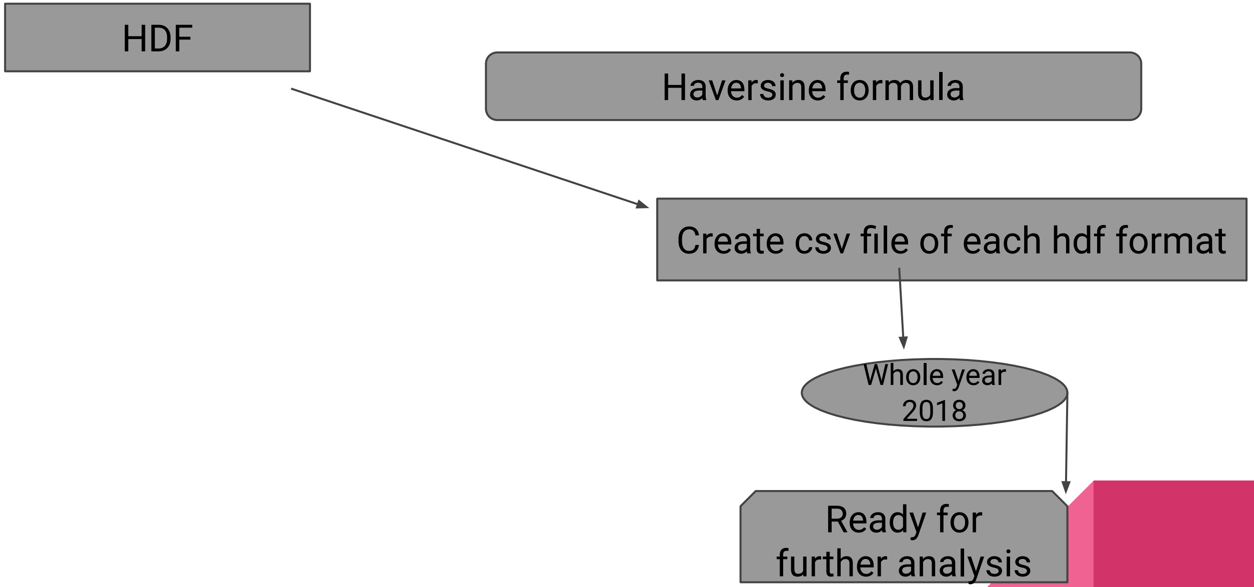
**Calculate Central Angle (c)**: Compute the central angle c between the user's location and each dataset point:

C = 2*np.arctan2(np.sqrt(a),np.sqrt(1-a))

d = R*c then find the nearest point and the aod value of this point is assigned to the user's latitude and user's longitude

# Collect Dataset

- From haversine formula we extract the Aod from hdf format



## Data Prepossessing

- ❑ Check Null values:
  - During the monsoon seasons (june  to september) many days aod values is missing
  - ❖ Step1 : so we interpolate this value by linear method

- ❑ Check this data is stationary?
  - Plot the time series plot of a particular place

# Linear interpolation Method

linear Interpolation

September 2023

## 1 Introduction

We've approached the interpolation problem by choosing high-degree polynomials for our basis functions $\phi_i$:

$$f(x) = \sum_{i=0}^{n} c_i \phi_i(x)$$

Recall the barycentric form of the Lagrange interpolant. However, using high-degree polynomials can lead to large errors due to erratic oscillations, especially near the interval endpoints. To mediate this, we'll try a different approach. We'll break up the interval over which the data is defined into small pieces, and we'll use a low-degree polynomial interpolant over each piece!

### 1.1 Piecewise Polynomial Interpolation

To begin, we'll consider the simplest case: piecewise linear interpolants (used by MATLAB when plotting).

To find this interpolant, we need only find the line between each pair of adjacent points on each interval:

$$s_i(x) = f(x_i) + m_i(x - x_i), \quad x_i \leq x \leq x_{i+1}$$

Here, $m_i$ represents the slopes in each interval. [Remark 0.0.1] On each subinterval $[x_i, x_{i+1}]$ for $i = 0, 1, \ldots, n-1$, the piecewise polynomial interpolant $s$ coincides with a linear polynomial, given by:

$$s(x) = s_i(x) = a_i + b_i(x - x_i),$$

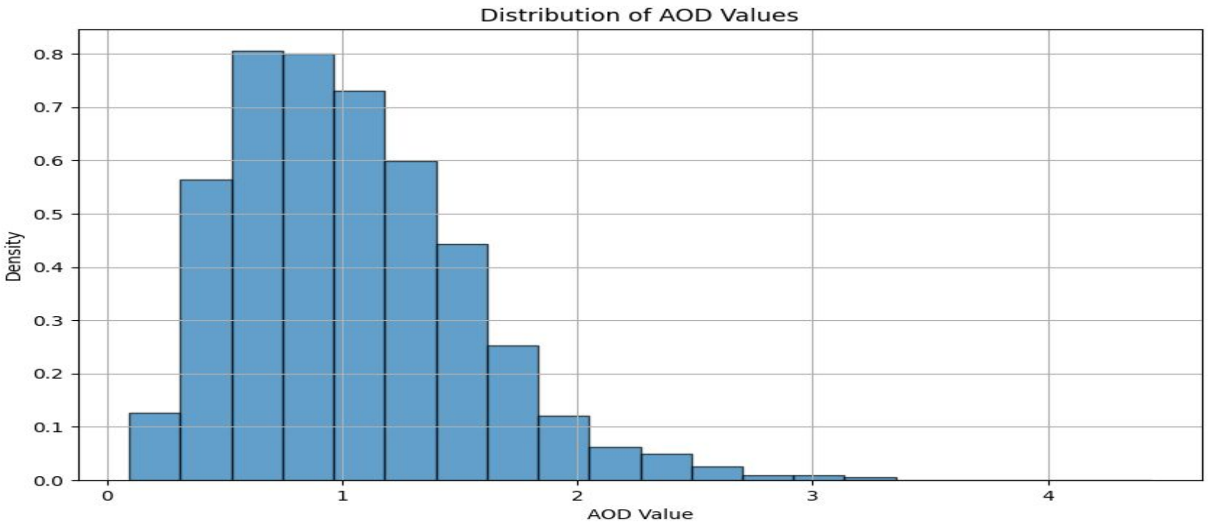
where:

$$a_i = f(x_i),$$

$$b_i = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}.$$

The values of $a_i$ are determined by the interpolation requirement $s(x_i) = s_i(x_i) = f(x_i)$, and the values of $b_i$ are determined by the requirement that $s$ be continuous, expressed as $s_i(x_{i+1}) = s_{i+1}(x_{i+1})$ for $i = 0, 1, \ldots, n-2$.

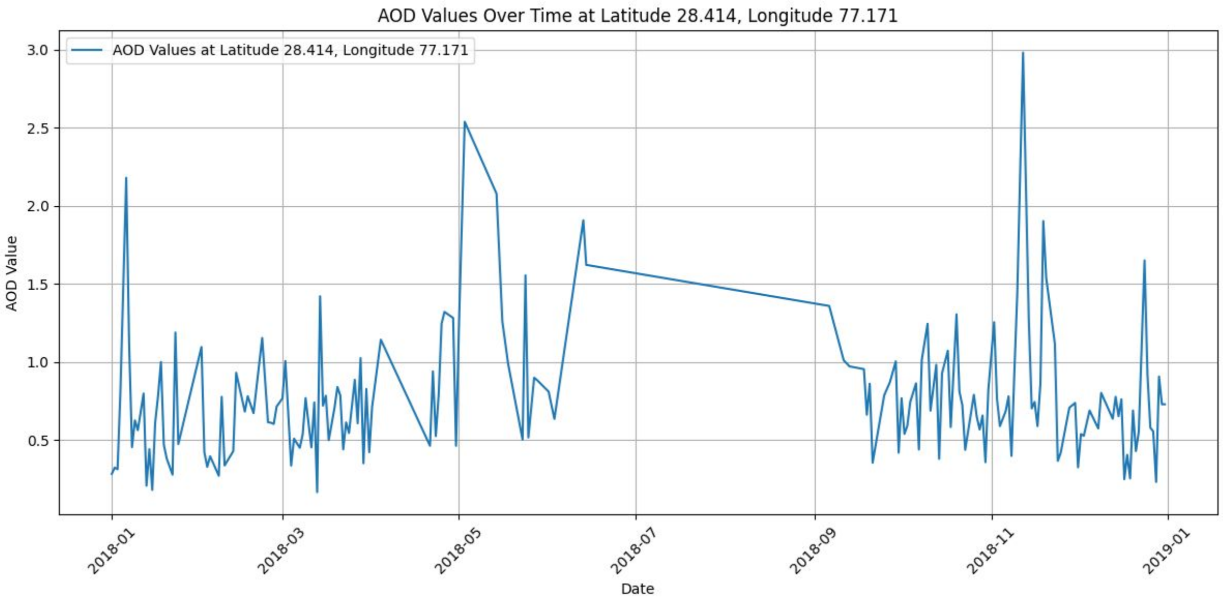
## AOD Distribution




From this plot we conclude that in the whole delhi region the maximum AOD value is above 3 and many place's AOD value is 0.25. Maximum number of places AOD value lies in the interval 0.50 to 1.75

In This Plot AOD distribution is right skew.

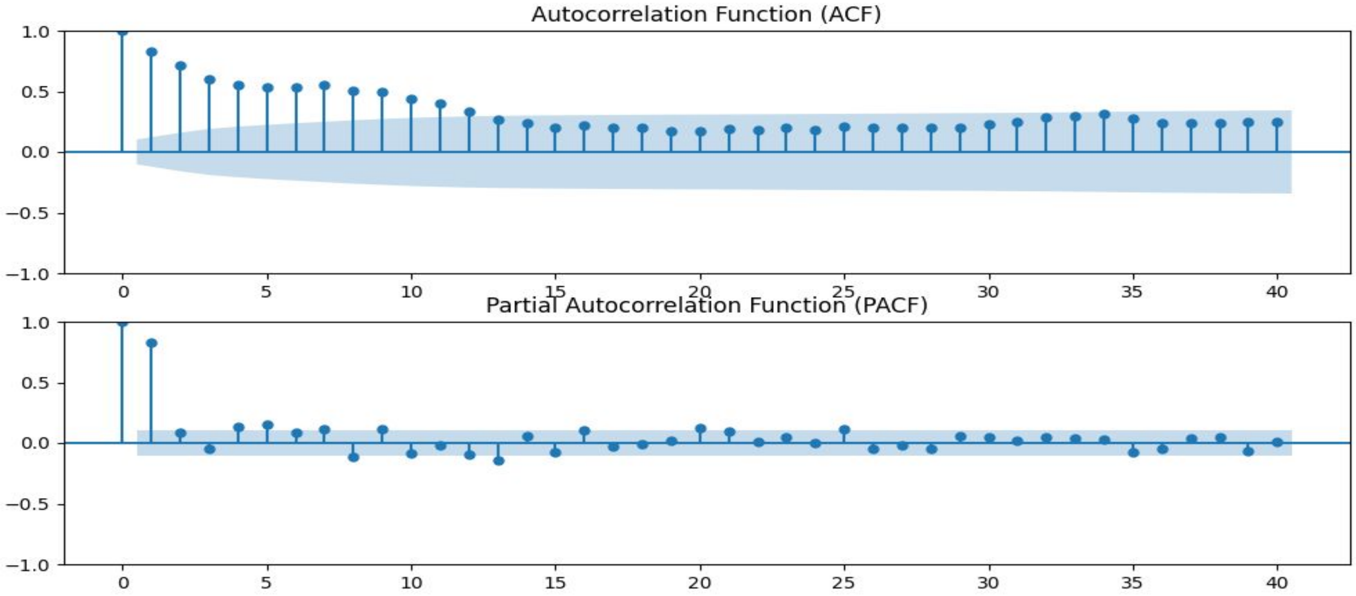
## Time series plot of the particular area




From this plot we conclude that the data is nonstationary. Because of the mean value is not constant in the whole timestamp

So I consider this data follows ARIMA model

# ACF and PACF plot of a particular latitude and longitude




We conclude that the data follows ARIMA(p,d,q)
Where p is 1 to 5 , q is 1 to 5 and d is 1 to 5

# BEST ARIMA Model

```
Ljung-Box Test Statistic: lb_stat
P-value: lb_pvalue
                               SARIMAX Results
==============================================================================
Dep. Variable:              aod_value   No. Observations:                  365
Model:                 ARIMA(4, 1, 4)   Log Likelihood                 -23.703
Date:                Wed, 13 Sep 2023   AIC                             65.407
Time:                        15:24:41   BIC                            100.481
Sample:                    01-01-2018   HQIC                            79.347
                         - 12-31-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3413      0.067     -5.056      0.000      -0.474      -0.209
ar.L2          0.9629      0.089     10.839      0.000       0.789       1.137
ar.L3         -0.0003      0.049     -0.006      0.995      -0.095       0.095
ar.L4         -0.8002      0.050    -15.870      0.000      -0.899      -0.701
ma.L1          0.1702      0.078      2.189      0.029       0.018       0.323
ma.L2         -1.1656      0.082    -14.268      0.000      -1.326      -1.005
ma.L3         -0.0488      0.070     -0.695      0.487      -0.186       0.089
ma.L4          0.7837      0.079      9.922      0.000       0.629       0.938
sigma2         0.0663      0.003     19.293      0.000       0.060       0.073
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):               169.71
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               0.89   Skew:                             0.64
Prob(H) (two-sided):                  0.52   Kurtosis:                         6.09
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

# Ljung Box Test

The Ljung–Box test may be defined as:

**$H_0$**: The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

**$H_a$**: The data are not independently distributed; they exhibit serial correlation.

The test statistic is:[2]

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

where *n* is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag *k*, and *h* is the number of lags being tested. Under $H_0$ the statistic Q asymptotically follows a $\chi^2_{(h)}$. For significance level α, the critical region for rejection of the hypothesis of randomness is:

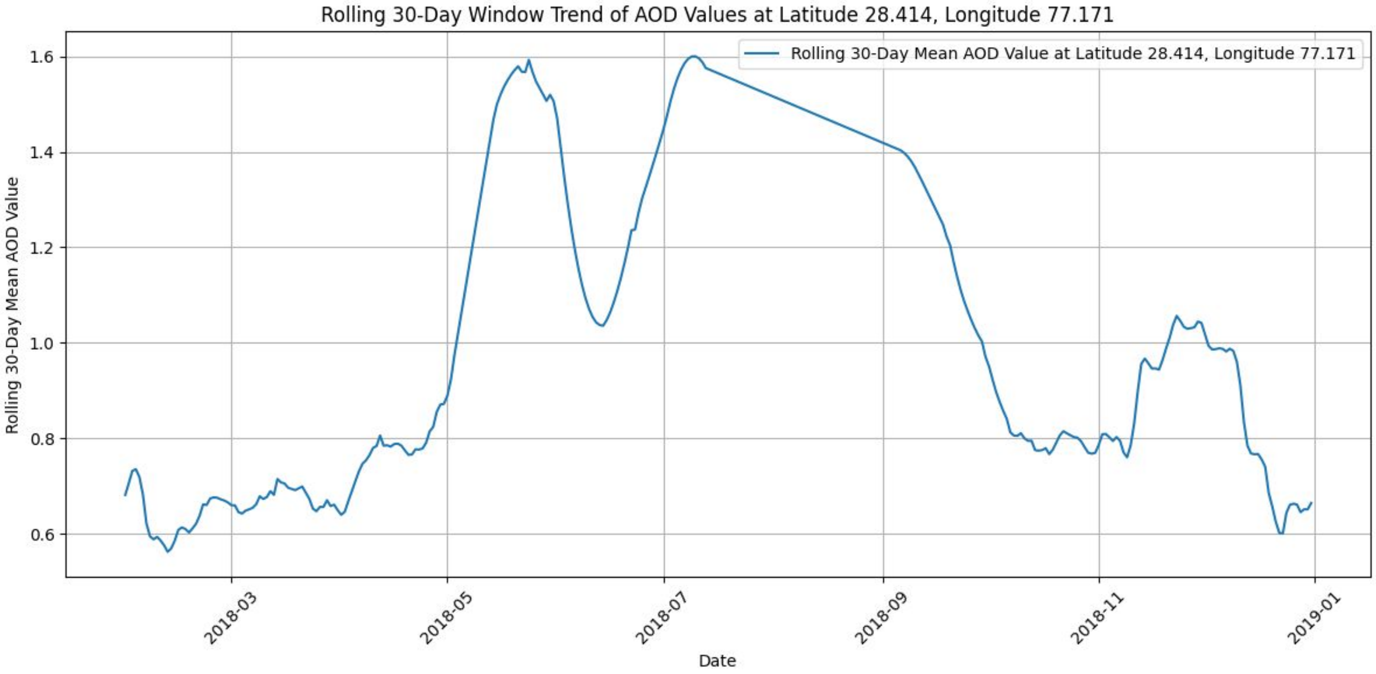
$$Q > \chi^2_{1-\alpha,h}$$

where $\chi^2_{1-\alpha,h}$ is the (1 − *α*)-quantile[4] of the chi-squared distribution with *h* degrees of freedom.

The Ljung–Box test is commonly used in autoregressive integrated moving average (ARIMA) modeling. Note that it is applied to the residuals of a fitted ARIMA model, not the original series, and in such applications the hypothesis actually being tested is that the residuals from the ARIMA model have no autocorrelation. When testing the residuals of an estimated ARIMA model, the degrees of freedom need to be adjusted to reflect the parameter estimation. For example, for an ARIMA(*p*,0,*q*) model, the degrees of freedom should be set to $h-p-q$.[5]

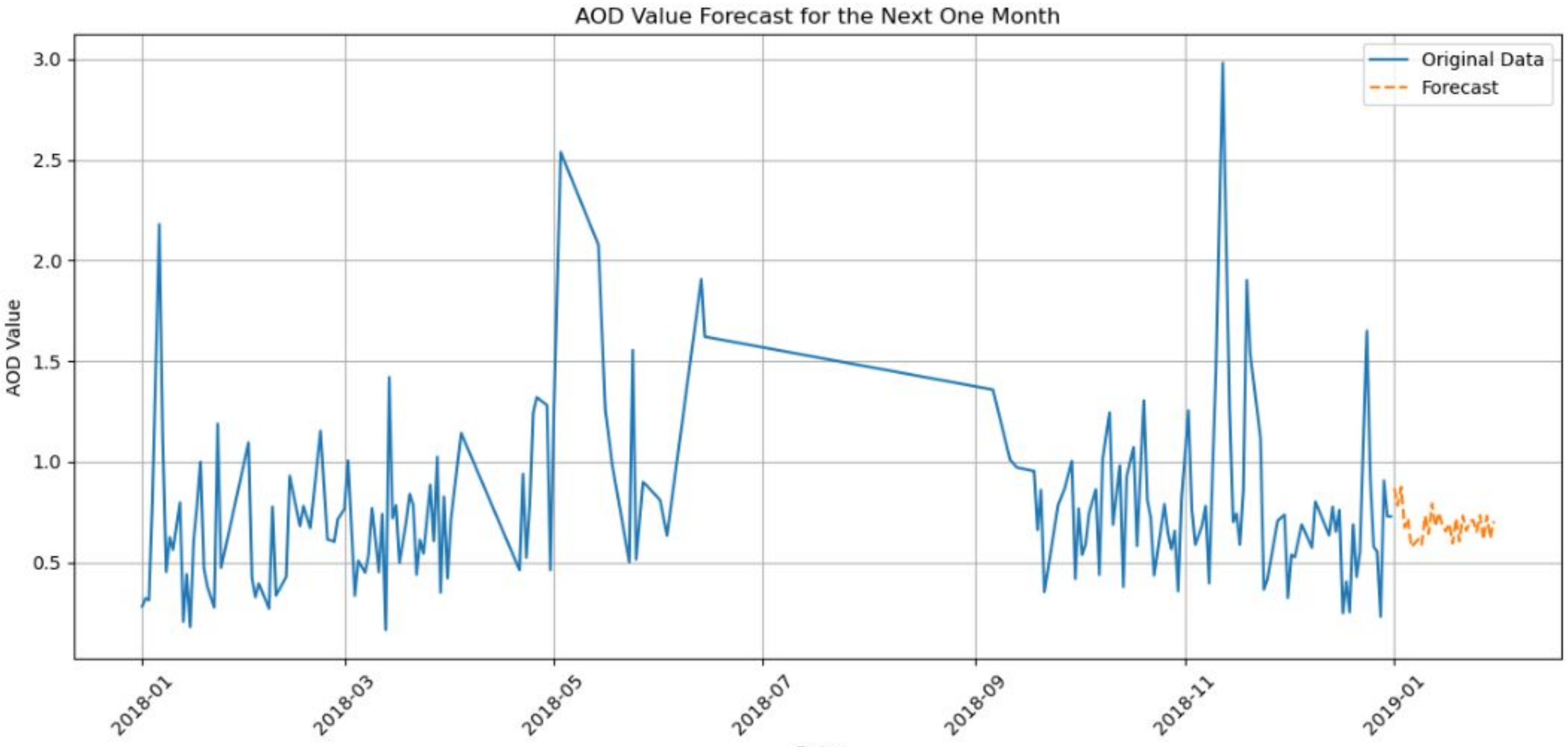
## Box–Pierce test

# 30 days windows trend AOD value plot

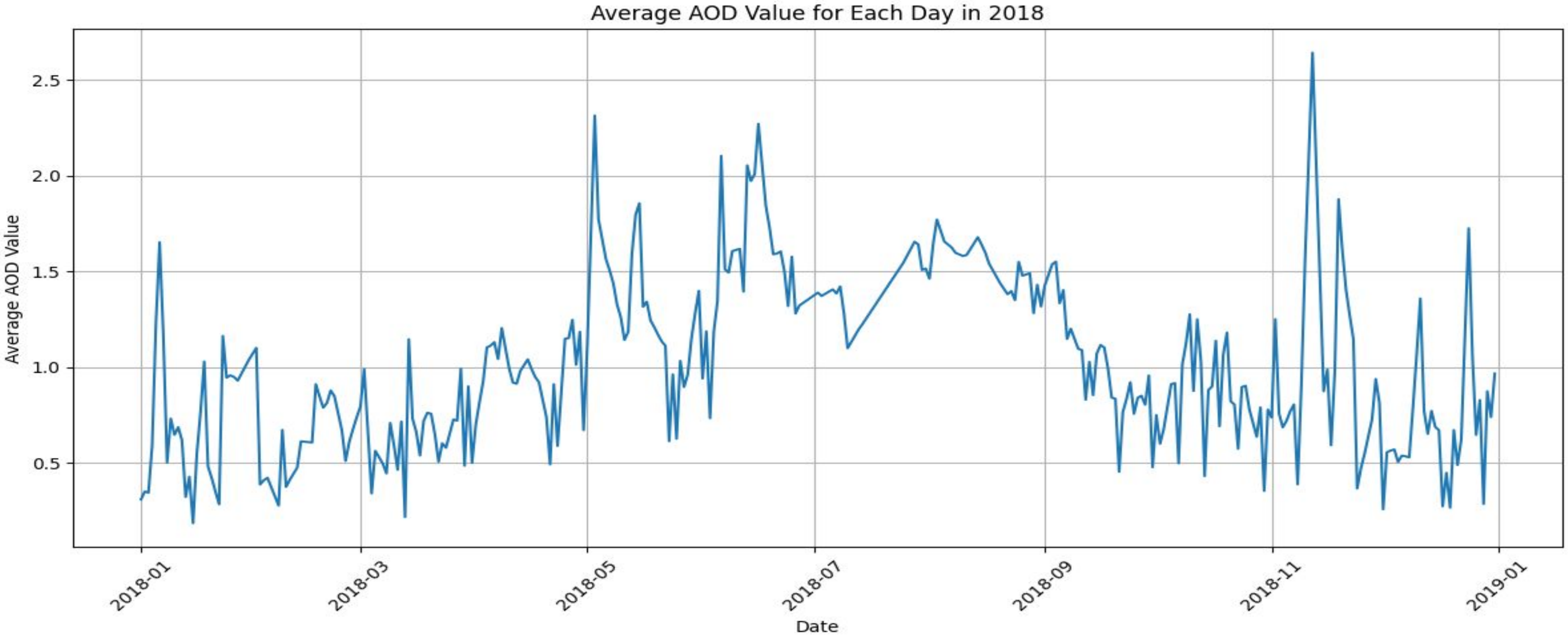



We conclude that 30 days windows average the AOD plot shows that in the month of june - july the AOD value decrease suddenly due to the monsoon reasons.

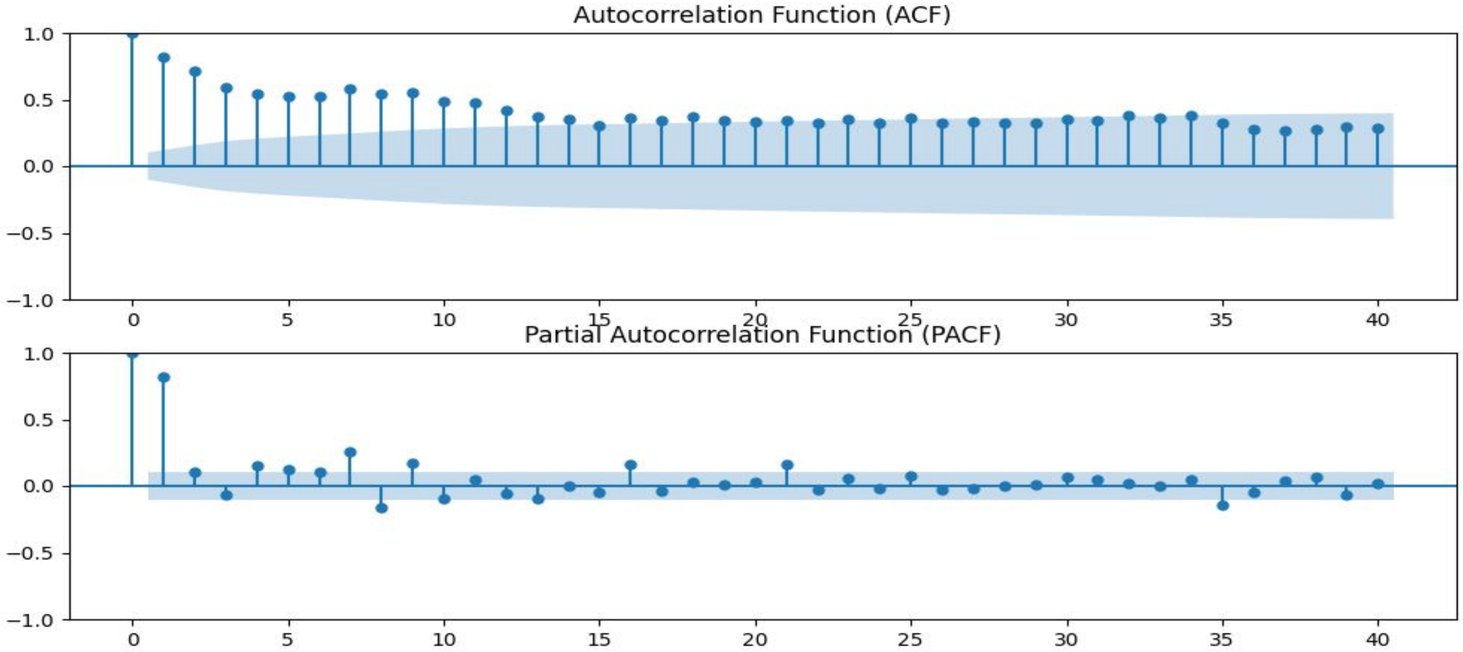
# Forecast one Month on the particular area



# Average AOD plot over year 2018



# MODEL : ARIMA(2,2,4)

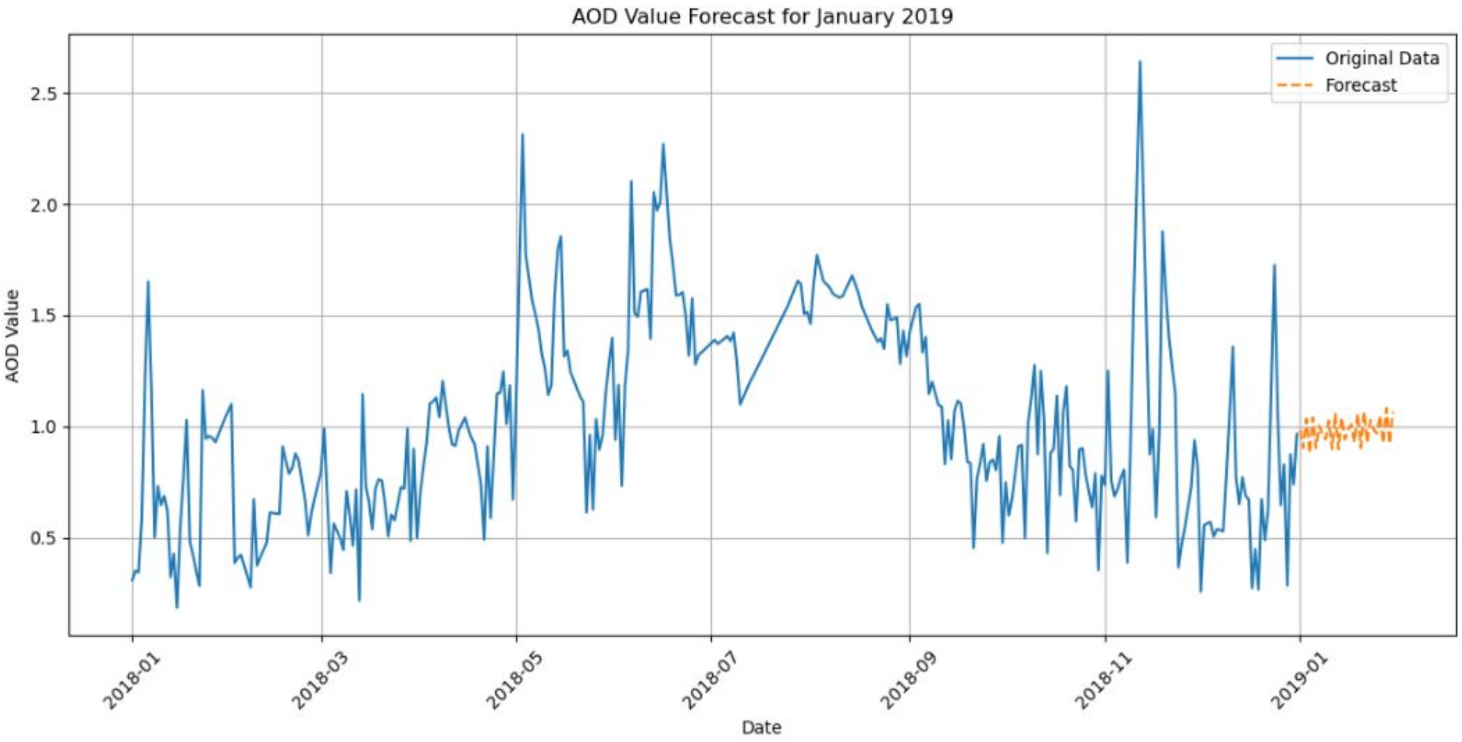


# BEST ARIMA MODEL

```
Best RMSE: 0.4965663572182941
Best Parameters (p, d, q): (2, 2, 4)
                               SARIMAX Results
==============================================================================
Dep. Variable:              aod_value   No. Observations:                  365
Model:                 ARIMA(2, 2, 4)   Log Likelihood                   2.352
Date:                Wed, 13 Sep 2023   AIC                              9.296
Time:                        15:49:10   BIC                             36.556
Sample:                    01-01-2018   HQIC                            20.132
                         - 12-31-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.8484      0.002  -1072.455      0.000      -1.852      -1.845
ar.L2         -0.9998      0.001  -1022.924      0.000      -1.002      -0.998
ma.L1          0.8023   2172.437      0.000      1.000   -4257.097    4258.701
ma.L2         -0.8788   3915.459     -0.000      1.000   -7675.038    7673.281
ma.L3         -0.9623   2006.361     -0.000      1.000   -3933.357    3931.432
ma.L4          0.0388     84.283      0.000      1.000    -165.152     165.230
sigma2         0.0557    120.910      0.000      1.000    -236.924     237.035
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               112.60
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               1.49   Skew:                             0.32
Prob(H) (two-sided):                  0.03   Kurtosis:                         5.65
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

## Forecast average AOD value of the january of 2019




ARIMA(2,2,4)

# Future Model

- ❏ Conv LSTM
- ❏ Vision Transformer

# Conclusion

- From this data analysis we conclude that the aod value increase in the time period of the diwali (november) and pre seasons of the Monsoon