Air Quality Assessment of Uttarakhand (India) using satellite data and machine learning techniques

Abstract

Degrading Air Quality is a major concern for all species on this planet. Over the years, it is seen that air quality is constantly degrading mainly of the reasons of industrialisation, deforestation, and green house effect. Main parameters to be considered with the Air Quality are the Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Ozone (O_3) and Aerosols. They are present in the air and their increasing or decreasing nature causes major changes in the air that organism's breath. A study of these parameters changing over time is necessary so to keep a check on the degrading air quality.

In this study, the data of Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Ozone (O_3) and Aerosols is taken for the past 5 years and their time series is extracted thereafter a test on stationarity is done so as to know whether these series are stationary or not. Two machine learning models namely Holt winter's Smoothing and FbProphet is applied to predict the value adjacent to the original value and a error metric is comparison is done to find out which model is best suited for forecasting these Air Quality parameters.

Keywords

Air Quality, FbProphet Model, Holt Winter's Method, Trend Analysis, Time Series, Time Series Analysis

Introduction

In recent years, with the acceleration of industrialization and modernization in cities, the problem of air pollution has become increasingly prominent, especially in areas with a large number of people which has seriously affected the daily production and life of residents. Air pollution is one of the most severe problems of the current time. It is growing day by day because of the vast level of industrialization and urbanization, causing massive damage to the flora and fauna of the planet. Every moment we are breathing air that is full of pollutants, going to our lungs, impregnating our blood, and then the whole body. causing uncountable health problems. Both state and central governments of India have put in many efforts to keep air pollution under control. Researchers performed a study over the Indian continent, the population is growing day by day and it is expected that soon India will become one of the

world's most populated countries[1].Other researchers proposed a study in which Seasonal and annual mean trends in aerosol optical depths (AODs) for the last decade are derived using MODerate Resolution Imaging Spectroradiometer (MODIS) Level 2 10 km *10 km remote sensing data over different locations in India. AODs have increased across India in the last decade [2]. Later conducted study to determine whether analytics models can be used to develop a system that can provide an approximation of future pollution levels with a wide degree of confidence. Techniques for rendered linear regression are found to be inadequate for the time-dependent data. In order to predict the future levels of various pollutants within a wide confidence range, we have used the time series forecasting approach. The effectiveness of our suggested method using SARIMA and Prophet model is shown by the experimental analysis of the forecasting for the air pollution levels of Bhubaneswar City[3].

It was seen in the past decade that the air quality of Uttarakhand has been degrading to quite an extent and proper measures need to be taken so that the state located in the lapse of Himalayas does not fall prey to industrialization and severely polluted air. Thus proper monitoring of air quality should be done with the help of satellite data and machine learning methods. Time series analysis of the satellite can be done to find the seasonality, trend, etc of the data and we can predict some future data, trends and seasonality on the air quality parameters (Time series analysis is a way of collecting and analyzing data points over some time.Some of the major pollutants responsible for degrading air quality are Carbon Monoxide (CO), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Aerosols, and Ozone (O_3). The Data of Air Quality Parameters has been taken from Google Earth Engine over the period 1th October 2018 to 30th December 2022. The Comma-Separated value file (CSV file) for this area is used to get the daily Air Quality Parameters. Many challenges need to be dealt with while making the dataset for the same like unavailability of data for the particular day, unavailability of the proper time step for the date index, etc.

Problem Statement

Degrading Air quality is quite a concern for the present, as well as the future coming generations so necessary actions, need to be taken so that it is not too late. So, necessary and accurate measures need to be taken so that proper monitoring of degrading air quality is taken, and not much. Quick, consistent, adaptable, cost-effective, and current information is the key challenge in air quality assessment. Some regulations must be made at the end of decision-makers so that proper monitoring and better policies and programs are made for regulating and regularly monitoring the air quality parameters. Some remote sensing

techniques can be employed for actual planning and creates degradation maps and find time series out of it. The final time series can be used to find trends, and seasonality and also to use machine learning algorithms to see how can we closely forecast corresponding to that time series with the least of errors.

Research Assumptions and Objectives

- Extracting Time Series of Air Quality Parameters from 2018 to 2022.
- Finding if the series is stationary or not.
- Finding Trend, Seasonality, ACF, PACF, Histogram, etc from that derived time series.
- Computing Air Quality Parameters with Holt Winter's and FbProphet Model.
- Comparing Holt Winter's and FbProphet with MAE, MSE, and RMSE.

Material and Methods

The methodological section of this project was broken down into two parts: the theoretical framework, which mainly focuses on research carried out and is mainly associated with the study's objectives as that we can choose objectives that seem more appropriate to the needs of the project. The second section illustrates how those approaches were used which we specified in the previously mentioned section. Primarily the approaches of the study include a process to obtain the data and analyze it and assess the data on various techniques for data visualization and mapping. These are described in the sections below.

Methods of Data Collection

Data is collected from the SENTINEL-5P satellite through Google Earth Engine and is processed in Google Collaboratory.

Data collected by the Sentinel-5 Precursor mission instrument is helpful for determining air quality. With a spatial resolution of 0.01 arc degrees, the TROPOMI instrument's multispectral sensor measures reflectance at wavelengths critical for determining atmospheric concentrations of ozone, methane, formaldehyde and carbon monoxide, nitrogen oxide, and sulphur dioxide as well as cloud characteristics. Data collected from the MODIS Terra satellite is used for finding the Aerosol Optical Depth in the Uttarakhand region of India. Also, the data is collected in the form of images first then it is converted into the form of Time series using the Google earth engine. While for getting the images we are using Google Earth Engine and for the latter part google Colab is used.Air Qualiuty Parameters which are

taken are Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Aerosol Optical Depth, Ozone (O_3), Carbon Monoxide (CO).

Study Area

According to the census of 2012, Uttarakhand had a population of 1.01 crores. In recent years the state has witnessed a rise in air pollution and a deteriorating air quality index and this is due to Carbon Monoxide (CO), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Aerosols, and Ozone (O_3) parameters. The Data of these parameters is Taken from Google Earth Engine. The Comma-Separated value file (CSV file) for this area is used to get the daily Air Quality parameters value.

Preprocessing data

The strategy includes first pre-processing the missing daily data by doing the average of the above and below values corresponding to the missing value. Make the date column into the proper format of DD/MM/YYYY format so that there is no ambiguity while processing it further. Also, many times data is multiplied by 10, 100, and 1000 to make it come in the range of 0 and 1.

In the case of using LSTM, data is first normalized between 0 and 1 so that proper LSTM can be induced in it. Some parameters such as Ozone, SO_2, contain a value negative so that value needs to be brought in the range of 0 and 1 so the normalization of data is required. As LSTM does not work well with negative values.

AD Fuller test

The ADF test is fundamentally a statistical significance test, which is another important thing to keep in mind. This indicates that a null and alternate hypothesis are used in the hypothesis test, and as a result, a test statistic is computed and p-values are reported. Whether a given series is stationary or not can be inferred from the test statistic and the p-value.

Through the ADfuller() function in statsmodels.tsa.stattools, the statsmodel package offers a trustworthy implementation of the ADF test. The following outputs are returned:

- A p-value
- The test statistic's value
- Amount of lags taken into account for the test
- cutoffs for critical values.

You reject the null hypothesis and conclude that the time series is stationary when the test statistic is less than the indicated critical value.

The number of lags you want to take into account when performing the OLS regression is an optional argument that the adfuller() function accepts. If the p-value obtained is less than significance level of 0.05 so we can reject the null hypothesis and take the series as stationary.

Holt Winter's Method

In the Holt-Winters method, we can do Double as well as Triple smoothing **(Chatfield, 1978)**. Exponential Smoothing refers to the use of an exponentially weighted moving average (EWMA) to smooth a time series.

The Holt-Winters method itself is a combination of three additional, much simpler smoothing method components:

- Simple Exponential Smoothing (SES): This type of smoothing makes the level of the time series an unchanging assumption. Therefore, it cannot be applied to series that include seasonality, trend, or both.
- Holt's Exponential Smoothing (HES): Holt's exponential smoothing allows for the inclusion of a trend component in time series data, elevating it above simple exponential smoothing. Seasonal data are still insurmountable for Holt's exponential smoothing.
- The addition of seasonality is finally possible thanks to Winter's Exponential Smoothing (WES), a Holt's Exponential Smoothing extension. The Holt-Winters method is known as Winter's exponential smoothing.

Due to the fact that the Holt-Winters method literally combines three smoothing techniques and stacks them on top of one another, it is frequently referred to as triple exponential smoothing.

FBProphet Model

The Facebook Core Data Science Team has released FbProphet, a potent time series analysis tool. For executing time series analytics and forecasting at scale, it is a straightforward and simple to use package.

Prophet is a method for predicting time series data that uses an additive model to fit non-linear trends with seasonality that occurs annually, weekly, daily, and on weekends as well as during holidays. Strongly seasonal time series and multiple seasons of historical data are ideal for it. Prophet typically manages outliers well and is robust to missing data and changes in the trend.

Training and Testing Data

The datasets are divided into two subsets for machine learning.The first subset, referred to as the training data, is a section of our actual dataset that is used to train a machine learning model. It trains our model in this way. The testing data refers to the other subset. Below, we'll go into more detail.Usually, training data is larger than testing data. This is because we want to provide the model with as much information as we can in order for it to identify and learn useful patterns. When our datasets's data are fed to a machine learning algorithm, the algorithm recognises patterns in the data and draws conclusions. We have divided our data in the ratio of 80:20 where 80% of the data is for training while 20% is for testing which is used to validate our model.

Error Metrics Used (Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE))

1- **Mean Squared Error (MSE)-** The error of the estimator or predictive model developed using the specified set of observations in the sample is represented by the Mean squared error (MSE).
2- **Mean Absolute Error (MAE)-** The average of all absolute errors is known as the Mean Absolute Error (MAE).
3- **Root Mean Squared Error (RMSE)-** Root Mean Square Error (RMSE) is the residuals' standard deviation (prediction errors). The distance between the data points and the regression line is measured by residuals, and the spread of these residuals is measured by RMSE.

Methodology

The following methodology has been adopted for the air quality assessment. First, we derive the data of Air Quality Parameters such as Sulphur dioxide (SO2), Nitrogen Dioxide (NO2), Ozone, Carbon Monoxide, and Aerosol from the Sentinel-5P satellite through Google Earth Engine. Then pre-processing of data to make it into a compatible form such as and filling in missing values. Hence, New derived data is obtained after this process. Now, this newly

derived data is passed through three methods namely Holt Winter's, and FbProphet model. Finally, the data is split in the ratio of 80:20 and, training and testing are done on these data by each method. Plotting of the forecasted and actual data is done next. To know the accuracy of the models, parameters such as Mean Absolute error, Mean Square error and Root Mean Square Error is calculated which gives us insights into the model's accuracy. Finally, the model with the least errors is considered better than others for forecasting purposes.

Results

AD Fuller test on Air Quality parameters

Before proceeding forward with our analysis of the air quality parameters, first of all, we need to find out even if our series is stationary or not. For that, we need to do the AD Fuller test on the series and if the value of p comes below 0.05 we can conclude that our series is stationary and hence we can proceed further with different-different tests.All the values of AD Fuller test on various time series was always less than 0.05 indicating that each time series is stationary and we can move further for analysis .

FbProphet Model on Air Quality Parameters

It was observed that the study model is predicting quite well. The black dots represent the actual data of Aerosol Optical Depth while the Blue line represents the forecasted value and the blue region around the forecasted value represents the yhat_lower for the lower boundary of forecasting while yhat_upper represents the upper boundary of the forecast. So the forecast will remain between these two lines. it was also observed that the aerosol optical depth time series is following Trend pattern and various individual trend line on yearly, weekly, and monthly basis

For Aerosol Optical Depth Various Error metrics for AOD are RMSE= 0.3812, Mean=0.4324 and MAE=0.2755.

For Carbon Monoxide time Series error metrics is RMSE=0.0568 , Mean= 0.2703 and MAE=0.0407.

For Sulphur Dioxide time series error metrics is RMSE=0.1066 , Mean=0.6829 and MAE=0.0851.

For Ozone time series error metrics is RMSE=0.0064, Mean=0.1255, MAE=0.0050

For Sulphur Dioxide time series error metrics is RMSE=0.0573, Mean=0.0291 and MAE=0.0383.

Holt Winter's Method on Air Quality Parameters

In this section, Holt Winter's classical Smoothing model is applied to the Various Air Quality Parameters. Smoothing such as Single Exponential Smoothing, Double Exponential Smoothing and Triple Smoothing is Applied to the data and the best model is chosen to compute and forecast the result corresponding to our actual data. Holt Winter's Smoothing on Aerosol Optical Depth Time Series Single Exponential Smoothing, Double Exponential Smoothing and Triple Exponential Smoothing is applied to the Aerosol Optical Depth time Series and results are obtained and checked against parameters such as Mean Absolute error, Root Mean Square Error, Mean squared error.

For Aerosol Optical Depth Triple exponential smoothing was found to be the best for fitting the data and smoothing so the error metrics for this model is MAE=0.3511, MSe=0.1912, RMSE=0.4372 and Mean=0.4437.

For Carbon Monoxide time series triple exponential smoothing was found to be the best and the error metrics calculated was RMSE=0.00531, Mean=0.2754, MAE=0.0383, MSE=0.028.

For Nitrogen Dioxide time series triple exponential smoothing was found better than other smoothing techniques and the error metrics calculated durng this is RMSE=0.0531,Mean=0.2754, MAE=0.0383, MSE=0.00282.

For Ozone time Series , error metrics was MSE=0.000606, MAE=0.0066, Mean=0.129, RMSE=0.0081.

For Sulphur Dioxide time series, error metrics obtained was MSE=1.8276, MAE=1.1074, Mean=0.294 and RMSE=1.351.

Discussion

The present study concluded that Deep learning and machine learning models are accurate for predicting along Air Quality Components. These models are capable enough to predict daily data of these air quality parameters.The models were trained on 80% data and testing was done on 20% data.The model accuracies were compared using standard stastical measures.In the study it was observed that when the country went on lockdown in the year 2020 a drastic improvement was seen in the air quality and mainly that NO_2. SO_2, CO, AOD decreased in

this period and O_3 increased which showed that Ozone layer was healing at this period. But as soon as lockdown was opened, with the opening of industries and traffic, again degradation in air quality was observed,

Some of the conclusions are:

1. IF we see error by terms of Mean Absolute error then:
- In case of Aerosol optical depth FbProphet outperformed classical Holt Winter's model.
- In case of carbon monoxide again Holt Winter's model outperformed Fbprophet model.
- In case of Nitrogen Dioxide, Holt Winter's Model was giving better results than FbProphet.
- In case of Ozone, Fbprophet was better when compared with Holt Winter's.
- In case of Sulphur Dioxide, FbProphet was definitely better than HW.

2. If we see Mean Squared Error as a metric then:
- In case of Aerosol optical depth FbProphet outperformed classical Holt Winter's model.
- In case of carbon monoxide again FBProphet model outperformed HW model.
- In case of Nitrogen Dioxide, Holt Winter's Model was giving better results than fbProphet.
- In case of Ozone, HW was better when compared with FbProphet.
- In case of Sulphur Dioxide, FbProphet was definitely better than HW.

3. If we see Root Mean Squared Error as a metric then:
- In case of Aerosol optical depth FbProphet outperformed classical Holt Winter's model.
- In case of carbon monoxide again HW model outperformed FBProphet.
- In case of Nitrogen Dioxide, Holt Winter's Model was giving better results than fbProphet.
- In case of Ozone, HW was better when compared with FbProphet.
- In case of Sulphur Dioxide, FbProphet was definitely better than HW.

It was concluded that FbProphet Model worked well in most of the cases so it can be taken as more superior model when in comparison with Holt Winter's method. It showed excellent

results in Simulation. With the degrading air quality FbPRophet model can be used effectively to predict future predictions of that major pollutant.

Conclusion

Degrading Air Quality is surely a major concern for each organism living on this planet and great concern should be given to this both by people and government. Government is trying to regulate the degrading air Quality with many measures such as by persuading citizens to switch to electric powered vehicle, do more afforestation, etc.

The use of FbProphet model in this present study sincerely demonstrated that it can be effectively used to predict the results with more accuracy. The accuracy of the model can be increased with experimenting more with the dataset which is fed into the model more outlier detection can be done, less peaks are encountered while we are predicting.

References

[1] **Abish, B., & Mohanakumar, K. 2013**. A stochastic model for predicting aerosol optical depth over the north Indian region. *Int. J. Remote Sens.*, *34*(4): 1449-1458.

[2] **Basistha, A., Arya, D. S., & Goel, N. K. 2008.** Spatial distribution of rainfall in Indian Himalayas–a case study of Uttarakhand region. *Water Resour. Manag.* ,*22*(10), 1325-1346.

[3] **Chitranshi, S., Sharma, S. P., & Dey, S. 2015.** Satellite-based estimates of outdoor particulate pollution (PM10) for Agra City in northern India. *Air Qual Atmos Health*, *8*(1): 55-65.

[4] **Gao, M., Sherman, P., Song, S., Yu, Y., Wu, Z., & McElroy, M. B. 2019.** Seasonal prediction of Indian wintertime aerosol pollution using the ocean memory effect. *Sci. Adv*, *5*(7):41-57.

[5] **Gupta, P., Remer, L. A., Patadia, F., Levy, R. C., & Christopher, S. A. (2020).** High-resolution gridded level 3 aerosol optical depth data from MODIS. *Remote Sens.*, *12*(17): 28-47.

[6] **Lamsal, P. G., & Singh, J. 2021.** Impact of COVID-19 on the Air quality over China and India using Long- Term (2009-2020) Multi-Satellite Data 2. *Air Qual Atmos Health, 8(6):156-168.*

[7] **Muthukumar, P., Nagrecha, K., Comer, D., Calvert, C. F., Amini, N., Holm, J., & Pourhomayoun, M. 2022.** PM2. 5 Air Pollution Prediction through Deep Learning Using Multisource Meteorological, Wildfire, and Heat Data. *Atm.*, *13*(5): 822-835.

[8] **Pope, R. J., Graham, A. M., Chipperfield, M. P., & Veefkind, J. P. 2019.** High resolution satellite observations give new view of UK air quality. *Wx.*, *74*(9): 316-320.

[9] **Sanghani, A., Bhatt, N., & Chauhan, N. C. 2016**. A review of soft computing techniques for time series forecasting. *Indian J. Sci. Technol.*, *9*(1): 1-5.

[10] **Yu, Y., Si, X., Hu, C. and Zhang, J. 2019.** A review of recurrent neural networks: Lstm cells and network architectures. Neural Comput., 31(7): 1235–1270.

Table 1 Final Error metrics for Aerosol Optical Depth

Aerosol Optical Depth			
Model	MAE	MSE	RMSE
Holt Winter's	0.3511	0.1912	0.4372
FbProphet	0.2755	0.1539	0.3812

Table 2 Final error metrics for Carbon Monoxide

Carbon Monoxide			
Model	MAE	MSE	RMSE
Holt Winter's	0.0383	0.028	0.00531
FbProphet	0.0407	0.0027	0.0568

Table 3 Final error metrics for Nitrogen Dioxide

Nitrogen Dioxide			
Model	MAE	MSE	RMSE
Holt Winter's	0.0069	0.000073	0.0085
FbProphet	0.0851	0.008	0.1066

Table 0 Final error metrics for Ozone

Ozone			
Model	MAE	MSE	RMSE
Holt Winter's	0.0066	0.000006	0.0081
FbProphet	0.0050	0.15	0.0064

Table 5 Final error metrics for Sulphur Dioxide

Sulphur Dioxide			
Model	MAE	MSE	RMSE
Holt Winter's	1.1074	1.8276	1.351
FbProphet	0.0291	1.61e-4	0.0573

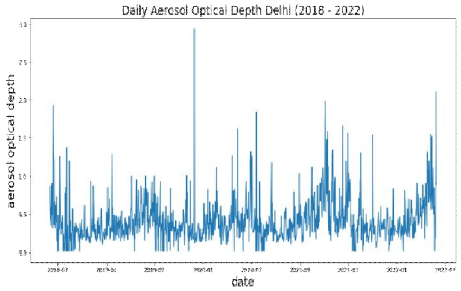


Fig. 1. Time series of carbon Monoxide

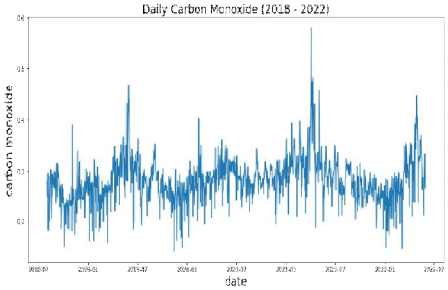
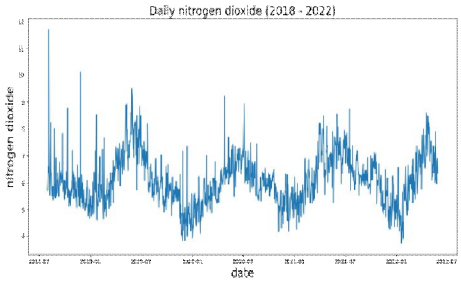


Fig. 2 Time series of Aerosol optical depth



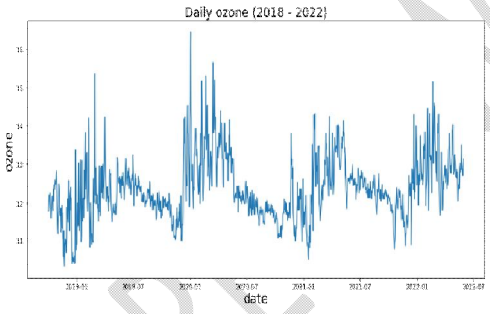


Fig. 3. Time Series of Ozone **Fig. 4. Time Series of Nitrogen Dioxide**

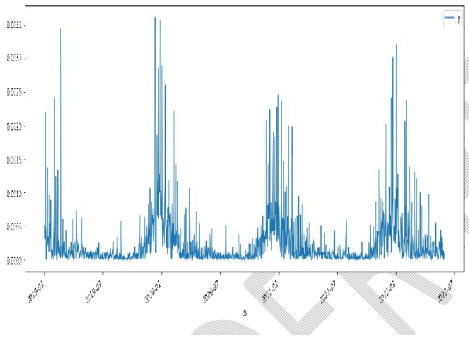


Fig. 5 time series of Sulphur dioxide

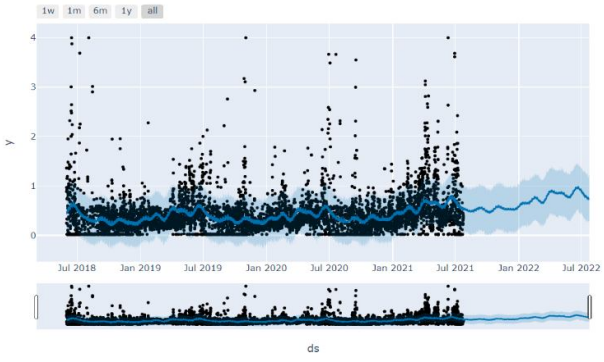


Fig. 6 FBProphet model on AOD

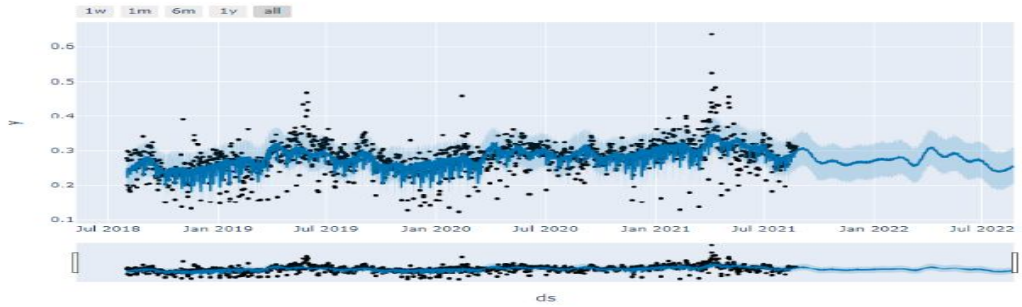


Fig. 7 FbProphet Model on Carbon Monoxide

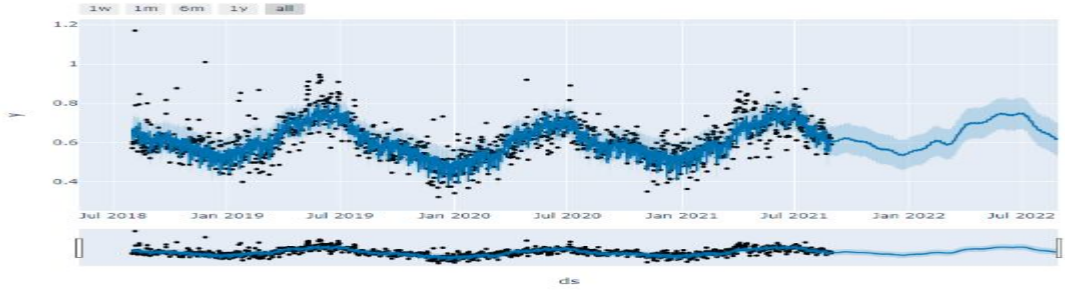


Fig. 8 FbProphet Model on Nitrogen Dioxide

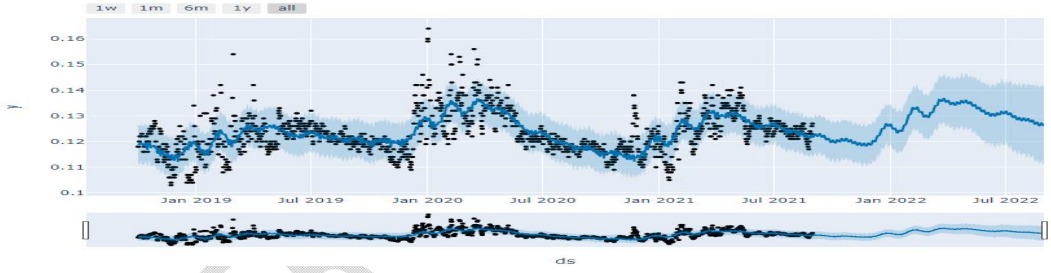


Fig. 9 FbProphet model on Ozone

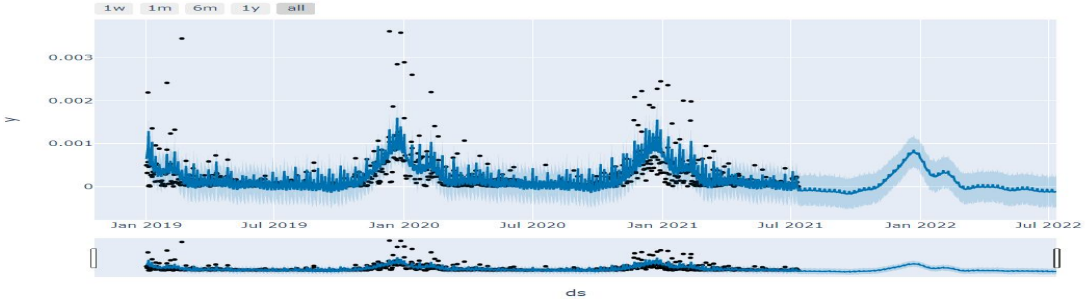


Fig. 10 FbProphet model on Sulphur Dioxide

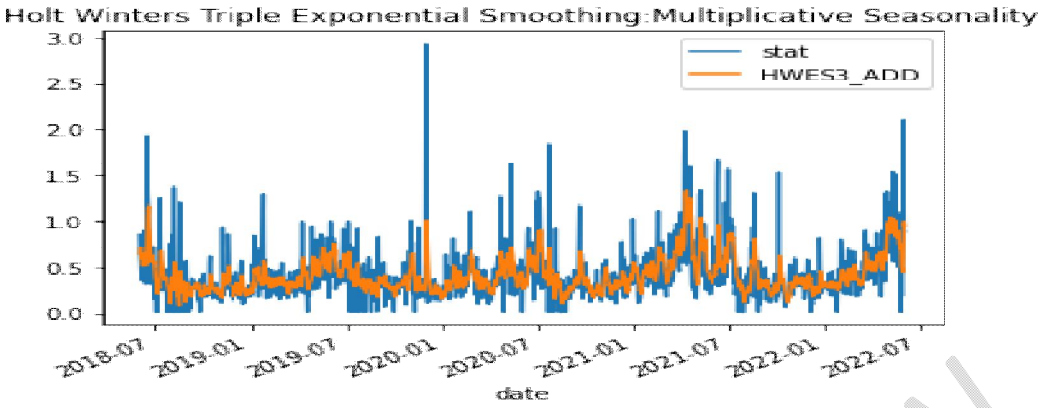


Fig. 11 Holt Winter's model on AOD

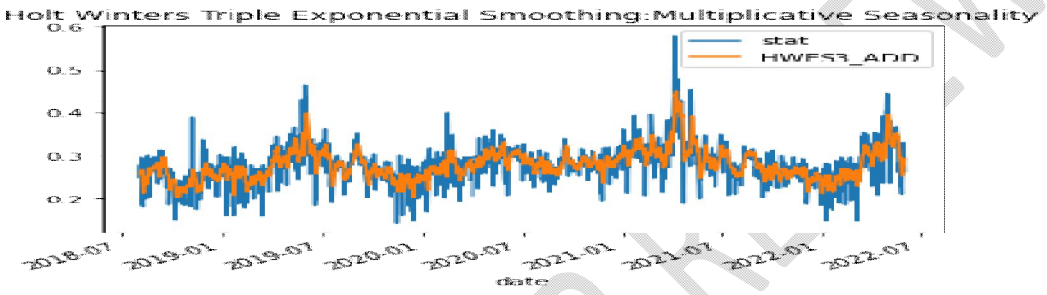


Fig. 12 Holt Winter's model on Carbon Monoxide

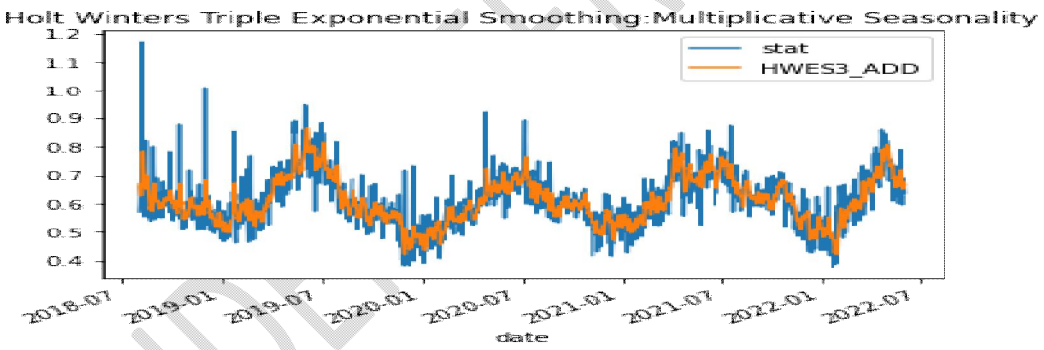


Fig. 13 Holt Winter's model on Nitrogen Dioxide

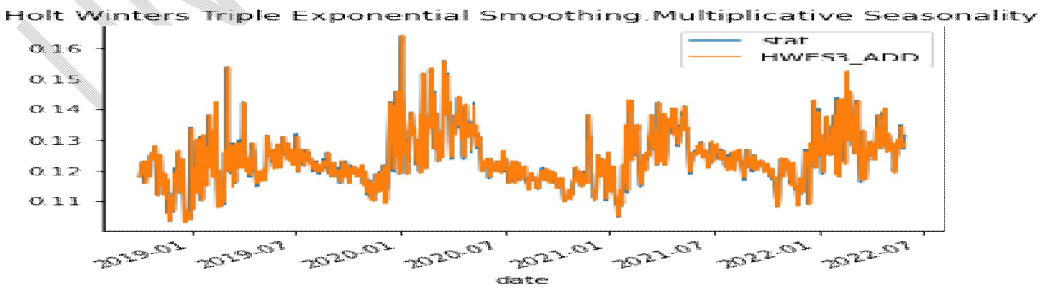


Fig. 14 Holt Winter's model on Ozone

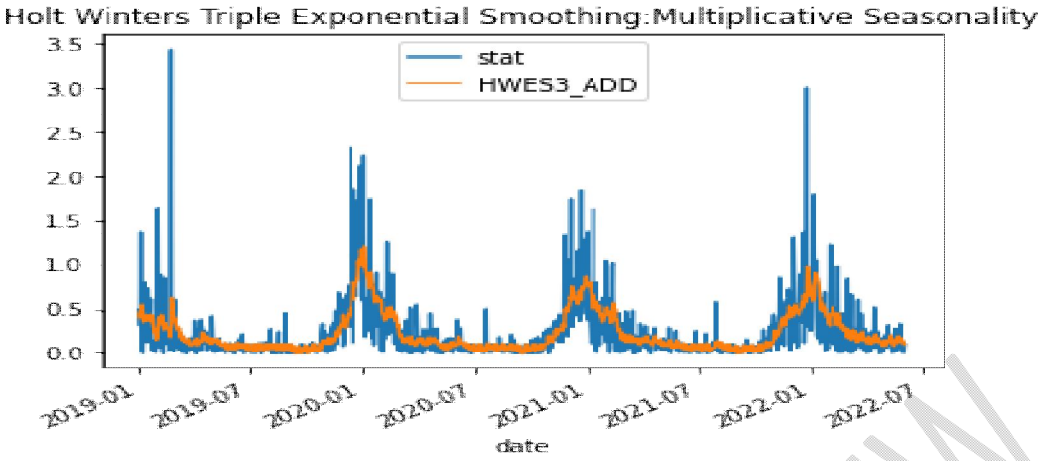


Fig. 15 Holt Winter's model on Sulphur Dioxide