Ministry of Earth Sciences

Indian Institute of Tropical Meteorology, Pune

Metropolitan Air Quality and Weather Forecasting Services

Forecasting of Daily Air Quality Index (AQI) based on the AI and Machine Learning models (ML) for Delhi metropolitan city

Introduction

1. Now a days, environmental crises such as hazardous waste, global warming, water and air pollution, resource depletion, and more have become significant concerns in the current and future era. Every year, millions of people die from diseases caused by air pollution. Public health is particularly vulnerable to the impact of air pollution in developing countries such as China and India.
2. Recently, AI and machine learning (ML) regression, neural and times series models gained increasing interests from investigators due to their superiorities compared with traditional statistical models, and had been applied for the prediction of air pollutants and/or air quality. The AI and ML models were good at solving non-linear relationship in which the variables from different sources had complex interactions.
3. Forecast of daily AQI is particularly challenging in metropolitan cities due to high population density, gas emissions, and uncertainty in meteorological processes.
4. In this research, we have used to eight independent variables such as 'PM10', 'PM2.5', 'NO', 'NO2', 'NOx', 'O3', 'CO and one dependent variables like AQI, which used to develop accurate models for daily AQI forecasting based on the these variables for Delhi City (Fig.1 to 3).

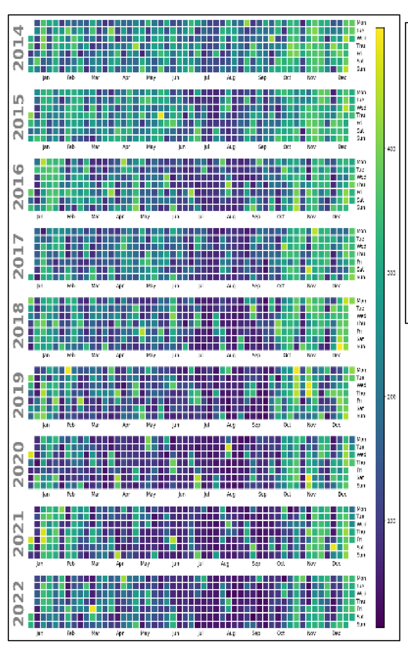


Fig.1 Calendar heat map of daily AQI datasets

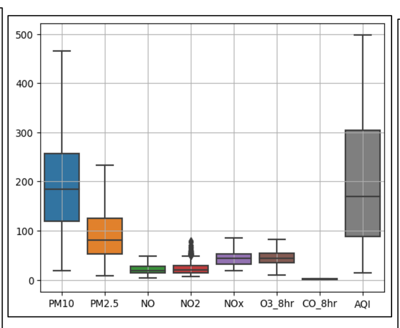


Fig.2 Box plot of Delhi city air pollutants datasets from 2014 to 2022

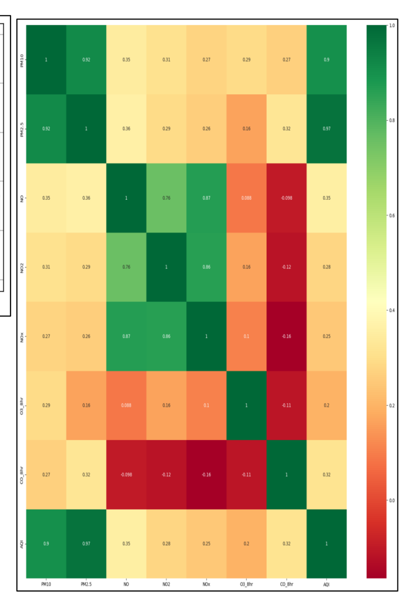


Fig.3 Correlation matrix of daily air pollution SAFAR datasets

Objectives

The main objectives of this study :

(1) To process the eight year SAFAR air pollutants daily datasets for calculation of AQI in Delhi city using python language.

2) Air pollutions datasets will remove errors before apply the AI and ML models, which is prepare various plots for understanding of the trends of air pollution and AQI in Delhi city.

(3) Development of the AI, Ml and times series models such as Linear Regression, Decision Tree Regressor, Random Forest Regression, SVR-Linear Kernel, XG-boost, Artificial Neural Network and ARIMA_ Time Series Models, which will be create the accurate model on the daily air quality datasets and python programming for forecasting of daily AQI.

(4) Compare AI and machine learning models aginst observations and to select the best model based on statistical metrics, which will be useful to accurately forecast daily AQI for Delhi city.

Methodology & Study Area

1. This study focuses on analyzing eight year air quality dataset for Delhi city. This research four main phases (Fig.4).
2. In the first phase, the eight year dataset is preprocessed to address any missing data and to gain insights through data visualization using different plots.
3. In the second phase, input variables are selected, and statistical analysis is performed to estimate the daily air quality index (AQI) values from 2014 to 2022 using Python packages.
4. In the third phase, five regression, one neural network and one times series time series models are developed to forecast daily AQI values for Delhi city.
5. The best forecasting models are identified based on statistical indicators, which can accurately forecast of daily AQI values for Delhi city.

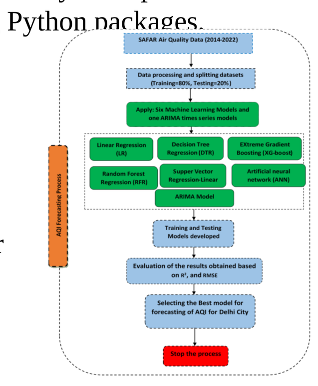


Fig.4 Flow chart of Methodology

Results

1. Air pollution is a major environmental issue that has been linked to numerous adverse health effects, including respiratory and heart diseases. Hence, there is a growing interest in developing accurate models for predicting air pollution levels in future.
2. Based on the performance metrics shown in the table 1, it can be concluded that the Decision Tree Regressor, Random Forest Regression, XG-boost, and Artificial Neural Network models have performed better than the Linear Regression and SVR-Linear Kernel models for the given dataset.
3. Among all the models, XG-boost has performed the best with the highest R^2 value of 0.99 and the lowest RMSE of 1.68 in the training dataset. It has also shown good performance in the testing dataset with an R^2 value of 0.99 and an RMSE of 4.69.
4. The Random Forest Regression model has also performed well with high R^2 values of 0.99 in both training and testing datasets and an RMSE of 1.18 in the training dataset and 4.72 in the testing dataset.
5. The Artificial Neural Network model has shown good performance with an R^2 value of 0.99 in the training dataset and an RMSE of 12.37. However, its performance in the testing dataset is slightly lower than XG-boost and Random Forest Regression with an R^2 value of 0.98 and an RMSE of 13.99.
6. The ARIMA Time Series Model has shown the lowest performance compared to the other models with an R^2 value of 0.84 in the training dataset and an RMSE of 41.23, and an R^2 value of 0.82 and an RMSE of 49.89 in the testing dataset (Fig.5a to Fig. 5g).

Table 1: AI and ML Model of Accuracy Performance for Daily AQI Values

ML Models	Training		Testing	
	R^2	RMSE	R^2	RMSE
Linear Regression	0.94	27.12	0.93	28.83
Decision Tree Regressor	1	0.01	0.99	6.12
Random Forest Regression	0.99	1.18	0.99	4.72
SVR-Linear Kernel	0.94	28.42	0.93	30.72
XG-boost	**0.99**	**1.68**	**0.99**	**4.69**
Artificial Neural Network	0.99	12.37	0.98	13.99
ARIMA_ Time Series Model	0.84	41.23	0.82	49.89

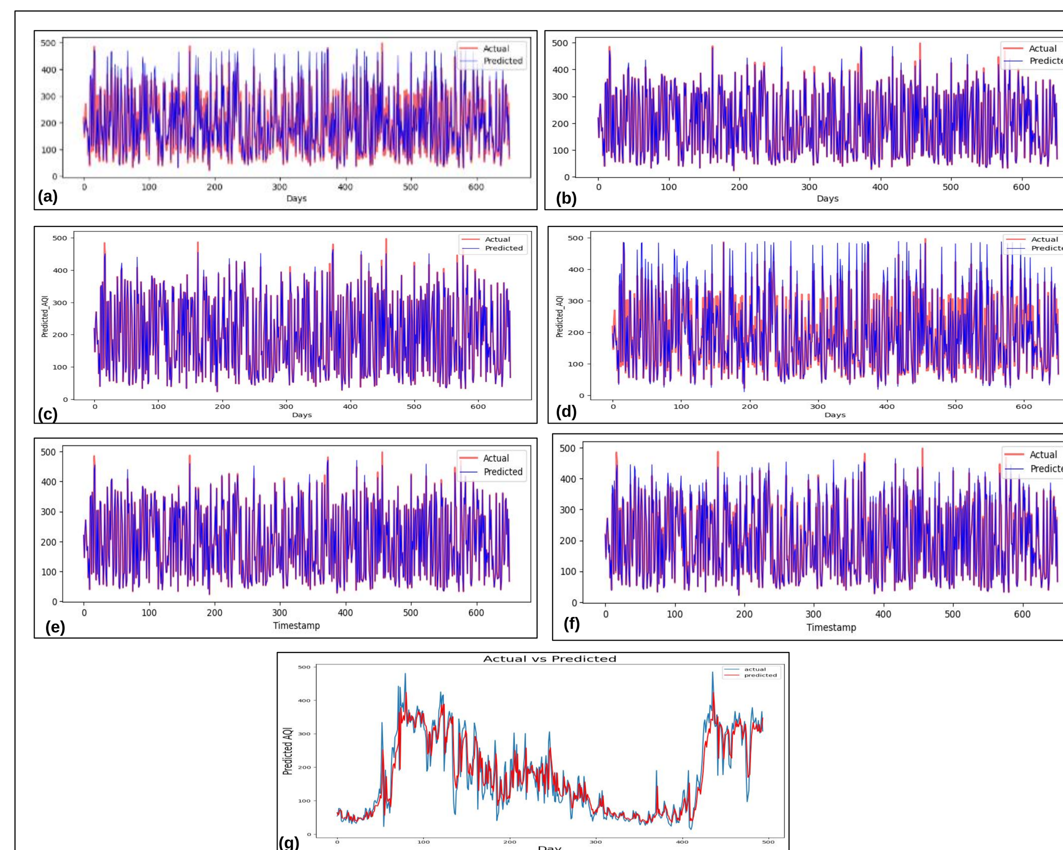


Fig. 5: (a) Linear Regression Model (b) Decision Tree Regressor Line (c) Random Forest Regression plot, (D) SVR-Linear Kernel, (e) XG-boost, (f) Artificial Neural Network (g) ARIMA model) times series model

Times Series Forecasting of AQI:

1). In this study, Based on the best model, we have forecasting the daily AQI values upcoming three years (Fig. 6).

2) The use of machine learning models, such as the ones evaluated in this study, could provide a valuable tool for predicting air pollutant levels and informing decision-making in this area.

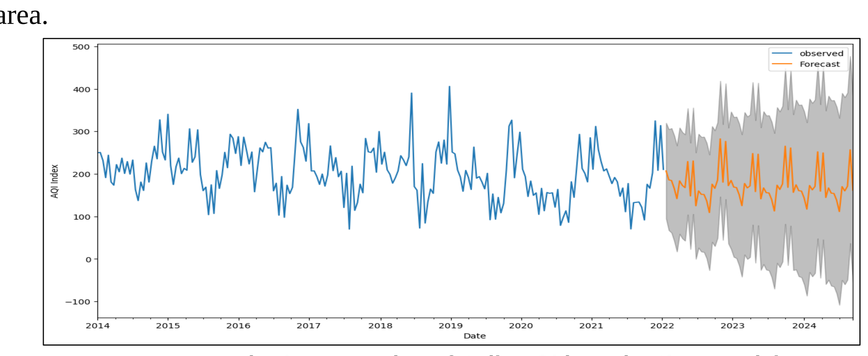


Fig. 6 Forecasting of Daily AQI based ARIMA model

Conclusion

1. Based on the findings and analysis, it is evident that air pollution is a significant environmental issue in Delhi, India, that poses severe health risks to individuals.
2. A set of seven models are validated for forecasting AQI values using air quality datasets. The models include Linear Regression, Decision Tree Regressor, Support Vector Regression with a linear kernel, XG-Boost Regressor, Random Forest Regressor, Artificial Neural Network models, and an ARIMA time series model. Each model has its strengths and weaknesses, and by comparing and evaluating the performance of each model, we can gain insights into the accuracy and reliability of the forecasts.
3. Therefore, it could be concluded that the XG-boost model has performed the best among all the models, followed by the Random Forest Regression and Artificial Neural Network models, for the given dataset.

Characteristics of atmospheric total gaseous mercury in an urban coastal city (Mumbai)

Introduction

1. Mercury (Hg) is an environmental toxic pollutant. It can impact both the human and environmental health due to emissions from different sources on the earth i.e. anthropogenic activities, ocean surface and fresh water bodies…etc.
2. Gaseous elemental Hg is most abundant ~95% of the atmospheric mercury, it has a long residence time (i.e. 0.5 to 2 yrs) in the air and is the most predominant form of Total gaseous mercury (TGM).
3. Coastal metropolitan cities are under the effect of Hg evasion from the sea surface and inland anthropogenic activities.
4. While pollution by atmospheric Hg is monitored in advanced countries, awareness is lacking in developing countries. However, the atmospheric Hg in the coastal metropolitan city of Mumbai has not been previously reported.
5. Therefore, continuous monitoring at metropolitan coastal cities in Mumbai is useful and it is necessary to understand how anthropogenic contamination, both present and future may have affected the atmospheric Hg in coastal urban areas.

Objectives

The prime objectives of present study:

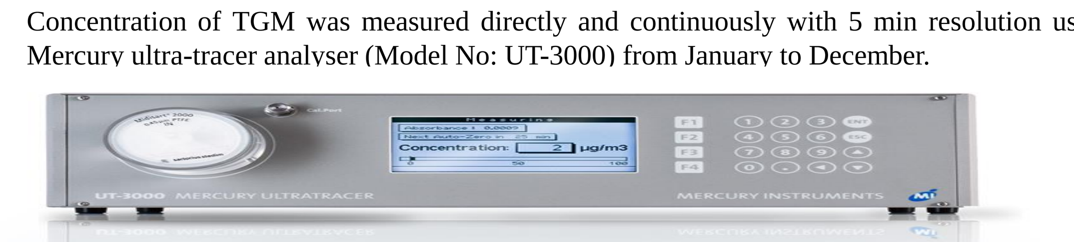
(i) To assess the trend of Hg contamination by comparing the concentration of TGM recorded during this study with previously reported data.

(ii) To highlight the influence of varying chemical pollutants and meteorological conditions on the atmospheric Hg distribution.

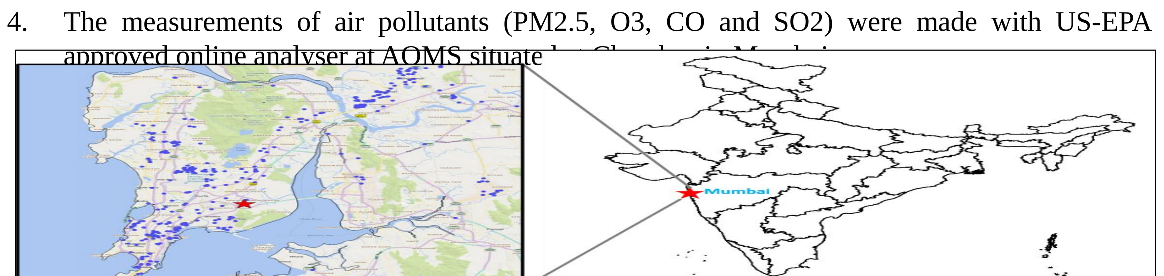
(iii) To identify the factors controlling changes in atmospheric mercury in the urban coastal area.

Methodology & Study Area

1. Concentration of TGM was measured directly and continuously with 5 min resolution using Mercury ultra-tracer analyser (Model No: UT-3000) from January to December.

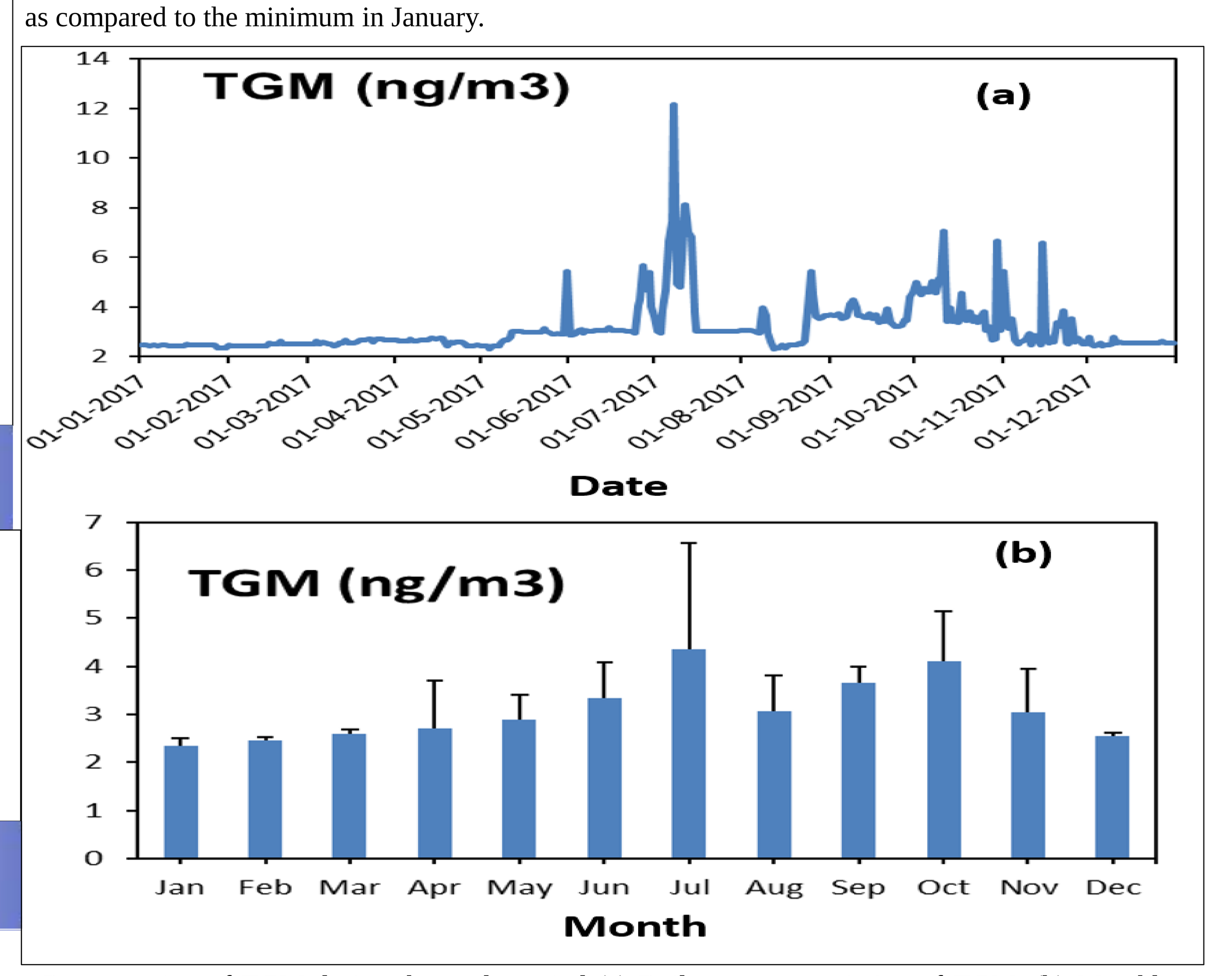


2. Briefly, analytical technique using Cold Vapor Atomic Fluorescence Spectrometry (CVAFS) is based on the collection of ambient mercury onto gold traps amalgamation followed by thermal deposition and final detection of Hg by CVAFS.
3. Meteorological parameters (temperature (T), wind speed (WS), wind direction (WD) and relative humidity (RH)) were measured simultaneously at the same sampling site with Automatic weather station (AWS).
4. The measurements of air pollutants (PM2.5, O3, CO and SO2) were made with US-EPA approved online analyser at AQMS situate

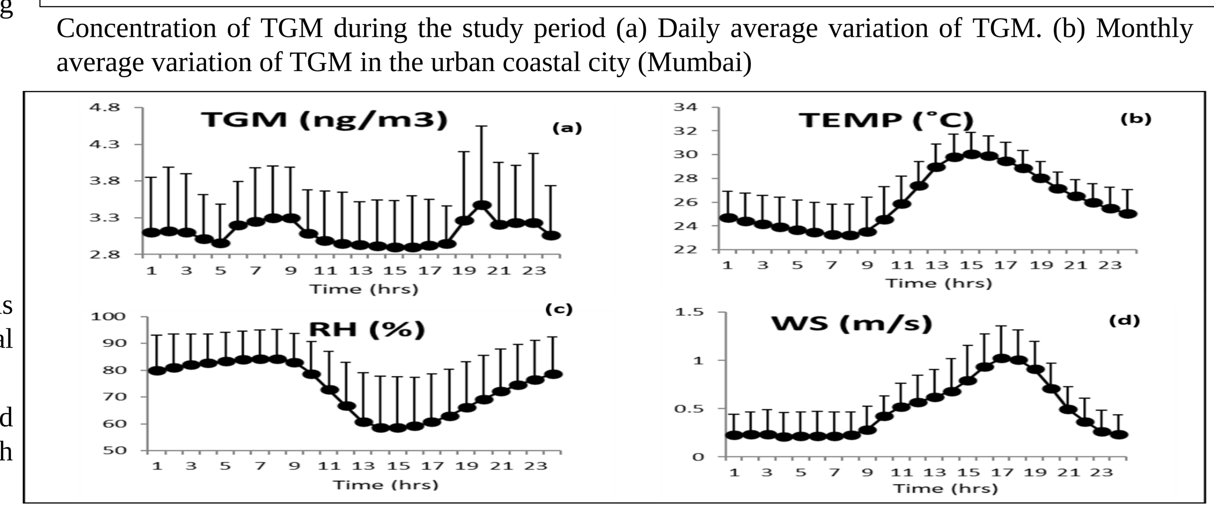


Results

The concentration of TGM ranged between 2.2 ng/m^3 and 12.3 ng/m^3 with an (avg. 3.1±1.1 ng/m^{-3}) which was significantly higher than the continental background values of around 1.5 ng/m^3 in the Northern Hemisphere. The highest concentration of TGM was observed in the month of July followed by October as compared to the minimum in January.



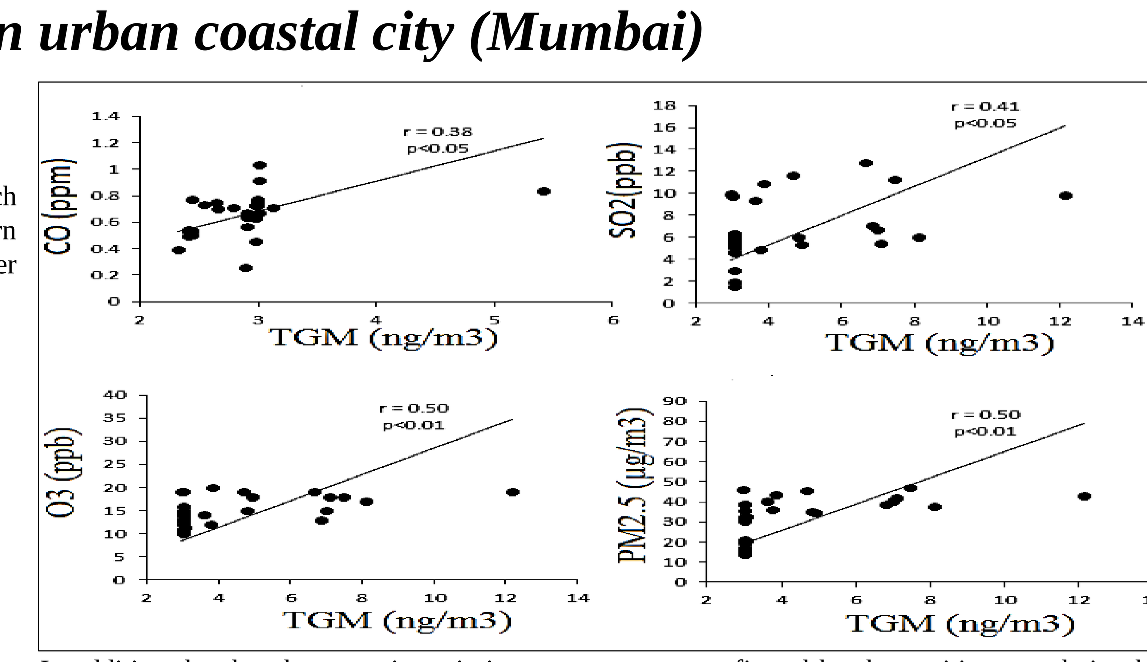
Concentration of TGM during the study period (a) Daily average variation of TGM. (b) Monthly average variation of TGM in the urban coastal city (Mumbai)



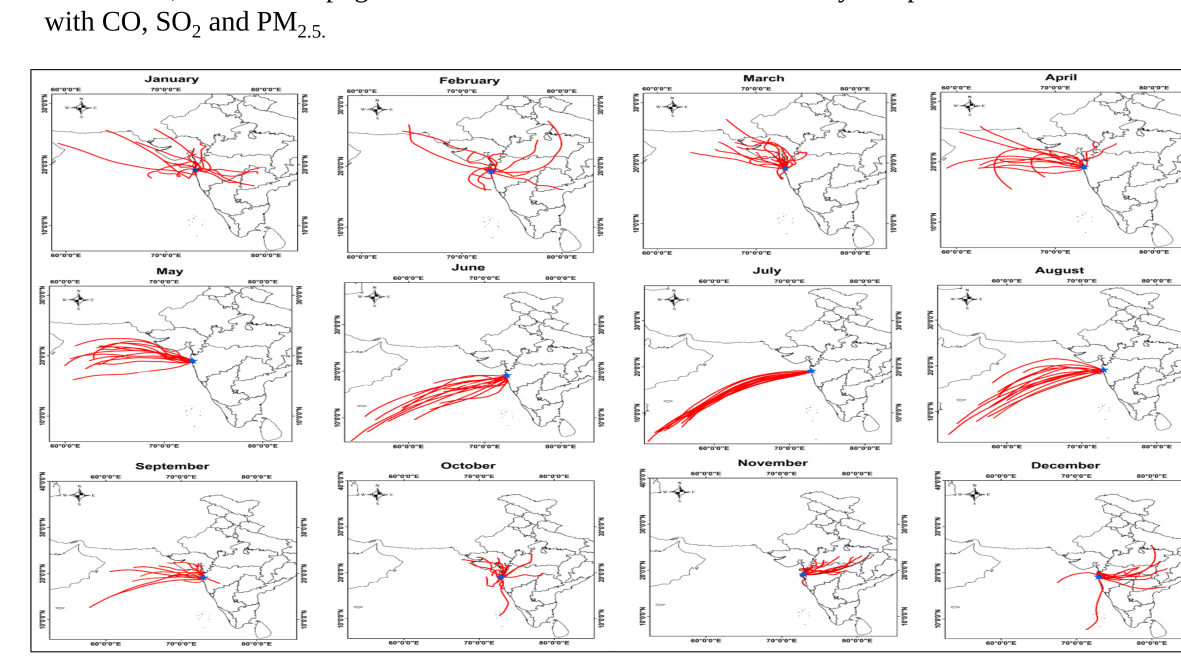
The average annual diurnal variation of TGM concentrations and meteorological parameters with standard deviation.

The concentration of TGM exhibited a distinct diurnal cycle mostly with an early night followed by early morning peak and an afternoon minimum.

The major peaks were observed during the early night and morning rush hours suggesting possible contribution from mobile combustions.



In addition, local anthropogenic emission sources were confirmed by the positive correlation between TGM with CO, SO_2 and $PM_{2.5}$.



Two days back trajectories were used to identify air mass origins during the study period (January to December) at the study site simulated by HYSPLIT vertical mixing model.

Conclusion

Signifying the influences of anthropogenic emission sources were confirmed by the positive correlation between TGM with SO2 and PM2.5 in July. Furthermore, positive correlation of TGM with temperature and relative humidity provides a strong indication that volatilization from the soil and sea surface are important in the study area.

Based on results of HYSPLIT backward trajectories, it indicated that the concentration of TGM was due to long-range transboundary air masses. The air mass was observed to come from the ocean and local anthropogenic emission sources during the study period.

Moreover, during the study period, the results of concentration of gaseous elemental Hg did not exceed the 200 ng/m3 recommended by the WHO as the regulatory limits for Hg in the atmospheric environment for long-term inhalation.