Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07616-4

ORIGINAL ARTICLE

# Hybrid learning model for spatio-temporal forecasting of $PM_{2.5}$ using aerosol optical depth

Pritthijit Nath[1] · Biparnak Roy[2] · Pratik Saha[3] · Asif Iqbal Middya[1] · Sarbani Roy[1]

Received: 2 December 2021 / Accepted: 4 July 2022


**Abstract**
Existence of several challenges and high cost in the development of monitoring infrastructure have become major reasons for data sparsity by statutory government agencies tasked to study pollution exposure in urban areas. As an effort to mitigate this problem, the recent usage of satellite aerosol optical depth data along with the usage of learning algorithms have become popular in recent times. This paper presents a novel four-staged approach using different machine learning, deep learning and statistical methods to develop a spatio-temporal hybrid model for temporal forecasting using data from existing stations along with satellite aerosol optical depth data for spatial interpolation. Experiments conducted on real-world data belonging to the cities of Kolkata, Bengaluru and Mumbai show that a consistent pattern is not followed in all the cities in all stages except in spatial interpolation where Random Forest Regression is found to surpass all other models used. While a long short-term memory network (LSTM Auto-Encoder) when employed in temporal forecasting inside the hybrid method outperforms others in Mumbai, a random forest regression-based method and a multi-layer perceptron-based method outperform others similarly in Kolkata and Bengaluru, respectively.

**Keywords** Air pollution · Hybrid models · Aerosol optical depth · Spatio-temporal modelling

## 1 Introduction

Exposure assessment [1] has gained valuable significance in the discipline of environmental science research for years with the importance increasing day by day due to continually increasing socio-economic activity. Effect of air pollutants on urban health has become one of the most important areas of focus particularly in recent studies worldwide. With increased awareness of the consequences of air pollution and its long-term impact on civilization, new fields of research such as urban computing [2] have gained huge traction recently. Governments all around the world have been involved in setting up monitoring stations in major urban centers to monitor the pollution scenario and draft environmental policies to help control rising pollution levels. As per a World Health Organization (WHO) report [3], more than four fifth of all people residing in urban regions have themselves vulnerable to air quality levels that exceed the prescribed WHO limits. Despite recent large-scale efforts in increasing monitoring stations [4], data sparsity, retrieval and processing still remain as one of the major engineering challenges that need to be studied in great detail, coupled with high real estate prices and expensive maintenance costs. Instrument failure also remains one of the most inevitable challenges agencies face when working in air quality analysis, as data due to faulty readings become missing, therefore requiring

✉ Sarbani Roy
sarbani.roy@jadavpuruniversity.in

Pritthijit Nath
pritthijit.nath@ieee.org

Biparnak Roy
roybiparnak@gmail.com

Pratik Saha
pratiksaha198@gmail.com

Asif Iqbal Middya
asifim.rs@jadavpuruniversity.in

[1] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

[2] Department of Instrumentation and Electronics Engineering, Jadavpur University, Kolkata, India

[3] Department of Computer Science, SRM University, Kattankulathur, Chennai, India

Published online: 29 July 2022

additional effort involving data imputation. Thus, not only temporal components, spatial components too pose a large enough challenge required to be tackled in great detail by researchers in this field.

$PM_{2.5}$, a type of particulate matter having diameter less than 2.5 micrometers, is one of the most widely studied contributors to air pollution. Sources of $PM_{2.5}$ include car engines, fuel based energy producing machines in households and industry, dust and soil particles, etc. These particles are of grave concern as they enter deep inside the respiratory system and are responsible for causing high medical damage [5] due to their tiny footprint. Exposure to excess $PM_{2.5}$ levels has been largely connected to cardiovascular and cerebrovascular diseases, neurological ailments, respiratory disorders, etc. [6]; thus, monitoring the concentration of $PM_{2.5}$ levels is of prime concern.

For monitoring and modelling of pollutants, in recent years, various spatio-temporal-based approaches have been used to perform spatial modelling and subsequent interpolation. To model the temporal aspects of the problem, Wang et al. [7] used the widely used seasonal autoregressive integrated moving average (SARIMA) to model $PM_{2.5}$ time-series data of long beach in Los Angeles. Lei et al. [8] used a hybrid approach to research the pollution effects in Fushung, Liaoning Province in China, where they used convolution neural networks (CNN) and Bi-directional long short-term memory networks (Bi-LSTM) to model the spatial features and predict temporal values, respectively. Wang et al. [9] used chemical transport models (CTMs) and land-use regression (LUR) models to investigate the empirical relationships between predictor variables and $PM_{2.5}$ actual values. While LURs were found to be restricted in their capacity to capture temporal variations, CTMs were often found to vary widely in their correspondence with pollution levels.

Long-term coverage and spatial seamlessness of data from several satellite-based instruments have become instrumental in monitoring of air pollutants especially in places where adequate monitoring stations have not been set up. AOD is a measure of the amount of prevention, atmospheric particles pose to sunlight in a column of air and is generally found to correlate empirically with surface $PM_{2.5}$ concentrations [10], thus making the use of AOD important in pollution-based studies.

Taking motivation from the recent advancements made to model both the spatial as well as the temporal aspects of the problem, this, a novel approach, is proposed in the paper to create a hybrid method to not only forecast the $PM_{2.5}$ values, but also spatially interpolate and produce predictions for the entire region in focus just from historical $PM_{2.5}$ time-series data from existing monitoring stations and AOD data

This study focusses on three key Indian cities of Kolkata, Bengaluru and Mumbai because of the air quality scenario and socio-economic factors prevailing in those cities. The pollution levels of Kolkata are around three-five times the permissible limit while Bengaluru is found to comparatively better than other cities . Mumbai is famous for the high socio-economic activity due to being the financial capital of India. The common aspect to the pollution scenario for all cities is the fact that the number of monitoring stations employed are not adequate to monitor every part of the city. In the novel four-staged approach proposed in this research article, various models from three different approaches namely statistical, deep learning and machine learning are used as temporal models along with different machine learning regressors as the intermediate models for interconversion of AOD to $PM_{2.5}$ and final spatial interpolation.

Compared to other works put forward by researchers, the four-staged approach proposed utilizes AOD due to high data availability through remote sensing as a key intermediate medium for spatial interpolation purposes. Using easy to implement baseline models which widely used for prediction tasks, the proposed hybrid approach combines the predictions made in various stages, to produce final predictions for unmeasured data points in the city grid being investigated. In doing so, this approach assists public pollution monitoring agencies in estimating pollution data at key places where monitoring stations are yet to set up.

To summarize, the main contributions put forward in this study are as follows:

- A novel hybrid approach is proposed to perform spatio-temporal forecasts from historical real-world $PM_{2.5}$ data derived from sparsely placed sites to predict the pollution values of unknown locations.
- Evaluation of various time-series models of different approaches including statistical, machine learning and deep learning is performed for finding the best model for temporal forecasts in the proposed hybrid approach.
- Usage of AOD as an intermediate medium to carry out spatial interpolation for $PM_{2.5}$ belonging to unmeasured city grid data points has been demonstrated.

The rest of the article is organized as follows: In Section 2, relevant studies conducted in this field are discussed. Section 3 consists of the problem formulation. Section 4 contains a detailed description of the framework, models used and the modelling approaches for the different stages of the proposed approach. The findings of this work as well as related discussion of the data are laid out in Section 5. The conclusion and impact of the study are ultimately presented in Section 6.

## 2 Related works

A survey of the recent related research studies conducted in the field of spatio-temporal pollution forecasting along with the use of AOD data is presented in this section.

An innovative method known as Multi-Scale Spatial Temporal Network (MSSTN) was proposed by Wu et al [17] for improved detection of multi-scale spatial and temporal patterns. MSSTN was used to make hourly predictions for $PM_{2.5}$ concentrations collectively for a number of cities on the basis of the training done on air pollution datasets of urban regions in North China. Lindström et al [18] proposed a method using spatial and spatio-temporal covariates. The data from several different monitoring stations were combined along with Geographic Information System (GIS) covariates to produce NOx predictions in the area around Los Angeles. To investigate the bias and variance in temporal modelling, Taieb and Atiya [19] conducted a Monte Carlo experimental study to analyze for multistep-ahead time-series forecasting and showed that a recursive one-step ahead strategy performs better for short time series and when there is a presence of a well specified model. To model chaotic time-series, Chandra [20] proposed a nature inspired cooperative coevolution (CC) optimization method for training recurrent neural networks. Xu et.al [21] used a Cubic spline method for series smoothing which later fed into an echo state network computed by an elastic-net algorithm. Soh et.al [22] used a combination of artificial neural network (ANN), CNNs, LSTMs was used to extract spatio-temporal relations for better performance. Zhu et.al [23] analyzed the causality relations present in air quality data using an efficient solution to overcome the computational complexity to process massive volumes of data along with adaptation of the grid-based technique used in different applications. An old study by Sahu et.al [24] implemented a fully Bayesian model using Markov Chain Monte Carlo techniques, which enabled the inference to process unknowns as well as perform predictions in space and time. In a related area under social network analysis, Cesario et. al [25] and Comito [26] utilized location-based social networks to understand human mobility and people behavior by mining check-in patterns, thereby studying the impact of structural patterns hidden in the nodes of a friendship network and external environment changes on the check-in patterns of the users.

Yang et al. [27] showed the lack of a detailed study into the association between AOD and $PM_{2.5}$. Their study investigated the relationships between AOD and $PM_{2.5}$ in 300+ cities of China within the time-period from 2013 to 2017, at various regional and temporal scales. Ni et al. [28] leveraged the backpropagation approach to model the $PM_{2.5}$ levels in the Beijing-Tianjin-Hebei (BTH) area during 2014 to 2016, in combination with the AOD and meteorological data. Mao et al. [29] used meteorological variables, AOD data and pollutant predictors as input and upon the use of geographic variability of $PM_{2.5}$ and AOD association was able to forecast pollutant concentrations 3 days in advance in eastern China. Kloog et al. [30] proposed the utilization of recent advances in Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data processing algorithms to use high resolution AOD data to generate grid predictions from generalized additive mixed models. Some notable related works to this study which inspired this study can be found mentioned in Table 1.

Unlike the existing AOD-based works discussed, the proposed hybrid approach differs in the use of the spatial interpolation based on AOD data and leverages the correlation existing between AOD and $PM_{2.5}$ found earlier to convert the interpolated values into their corresponding pollution values. This work utilizes various off-the-shelf methods from machine learning, deep learning and statistical approaches, which are widely available and can be implemented easily. The four-staged approach proposed, not only alleviates the need of good number of stations, but also enables in reducing costs by using only historical time-series pollution and freely available MODIS satellite data for processing the required results. In addition, the easy modularity of the inner-stages also helps in reuse in other related applications as well.

## 3 Problem statement

Given a set of $n$ known air quality monitoring stations $S = \{S_1, S_2, ..., S_n\}$ having pollution data $P_i = \{v_{i_1}, v_{i_2}, ..., v_{i_t}\}$ and AOD data $\Omega_i = \{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_t}\}$ for the $i$ th station over a duration of $t$ timesteps in the past, our intention is to find the pollution data $P_j = \{w_{j_{t+1}}, w_{j_{t+2}}, ..., w_{j_{t+t'}}\}$ over the next $t'$ timesteps for each $j$ th location belonging to the set of $m$ unknown locations $S' = \{S'_1, S'_2, ..., S'_m\}$. Here, $v$ and $w$ denote actual and predicted pollution values, respectively, while $\omega$ represent the AOD values extracted from satellite data.

## 4 Proposed approach

The approach to produce hybrid methods to tackle the problem formulated in this study can be divided into four folds. First, the pollution data for the $i$ th station, $P_i = \{w_{i_{t+1}}, w_{i_{t+2}}, ..., w_{i_{t+t'}}\}$ for station $S_i \in S = \{S_1, S_2, ..., S_n\}$ over the next $t'$ timesteps are found out using one-step ahead prediction method. For each $w_{i_t} \in P_i$, the

**Table 1** Notable mentions of relevant spatio-temporal pollution modelling research in the past few years

| Author | Year | Method | Description |
|---|---|---|---|
| Bui et.al [11] | 2020 | Spatio-Temporal Prediction using Multimodal Approach | Multimodal fusion of critical factors were used to predict future air quality levels thereby reducing the mean absolute errors of $PM_{2.5}$ prediction. |
| He et.al [12] | 2020 | Spatio-Temporal Neural Network (STNN) | STNN with an encoder-decoder architecture coupled with region-based spatial dependencies were used to improve accuracy. |
| Pu et.al [13] | 2020 | Spatio-Temporal modeling of $PM_{2.5}$ concentrations with missing data problem: A case study in Beijing, China | Built a multi-stage statistical technique in which daily $PM_{2.5}$ levels could be obtained with extensive spatial coverage. |
| Di et.al [14] | 2019 | An ensemble-based model of $PM_{2.5}$ levels over the contiguous United States with high spatio-temporal resolution | Developed an ensemble method which integrated multiple learning techniques and predictor variables to predict diurnal $PM_{2.5}$ at 1 sq. km resolution over the contiguous United States of America. |
| Stafoggia et.al [15] | 2019 | Estimation of daily PM10 and $PM_{2.5}$ concentrations in Italy, 2013-2015, using a spatio-temporal land-use random-forest model | Used ensemble models establish a relationship between measured PM and satellite, land use and meteorological parameters. Predictions were made over each 1 km x 1 km grid cell of Italy and were improved by using small-scale predictors computed within a small buffer area around the location. |
| Krishna et.al [16] | 2019 | Surface $PM_{2.5}$ Estimate Using Satellite Derived Aerosol Optical Depth over India | Integrated MODIS AOD data along with simulations from weather research and forecasting model with chemistry (WRF-Chem) model to ascertain the $PM_{2.5}$ levels at a resolution of 36 km over India. |

corresponding AOD value $\Lambda_{i_t}$ is then found out using a regressor. Based on the AOD predictions made, spatial interpolation methods using satellite AOD values $\Omega$ are then used to calculate the AOD values $\Lambda_{j_t}$ at each unknown location $S_j \in S' = \{S'_1, S'_2, ..., S'_m\}$. Finally, these interpolated AOD values are then converted back to calculate the pollution values $P_j = \{w_{j_{t+1}}, w_{j_{t+2}}, ..., w_{j_{t+t'}}\}$ for each $S_j$. This section deals with the framework involved coupled with the description of the modelling performed as a part of the study.

## 4.1 Framework

Detailed description of the various phases involved in the study is described as follows.

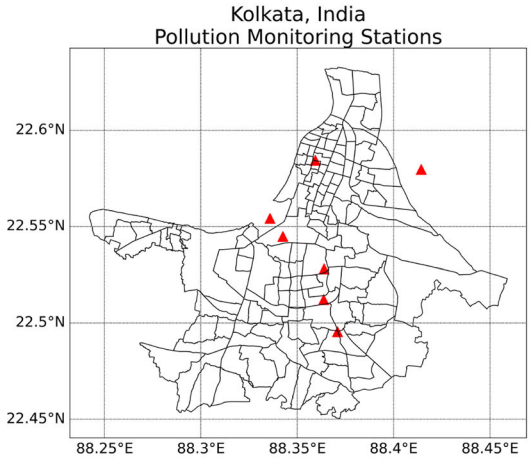
### 4.1.1 Data collection

The dataset used involves both pollution time-series data and AOD values over Kolkata, Bengaluru and Mumbai. The pollution time-series data containing daily $PM_{2.5}$ values for the monitoring stations depicted in Fig. 1e over the time-period of August 2018 to February 2020 had been extracted from the Kaggle dataset "Air Quality Data in India (2015-2020)" [31] which in turn compiled the data from the archives maintained by the Central Pollution Control Board (CPCB) [32] and other regional pollution monitoring stations. As the dataset did not contain the geographic coordinates, the geocoding of all the stations depicted in Fig. 1e was done using Google Maps Geocoding API.

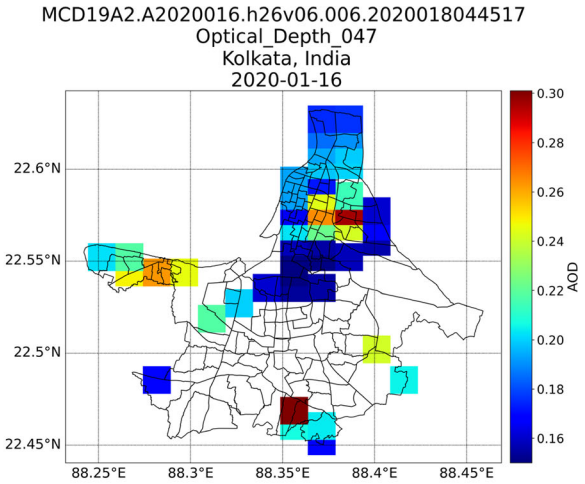
The AOD data used had been extracted from the MCD19A2 product data acquired from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument [33] aboard the TERA satellite made available through a web interface as a part of the Level-1 and Atmosphere Archive and Distribution System Distributed Active Archive Center (LAADS DAAC) [34]. Out of the 13 datasets provided in the MCD19A2 product, the blue band Optical Depth at 0.47 $\mu$m is used in the study. As the data were provided in tile format, where each tile contained AOD data corresponding to a section of the earth, tile having family ids h26v06, h25v07 and h24v07 associated with Kolkata, Bengaluru and Mumbai as shown with examples in Fig. 1b, 1d and 1f were used.

### 4.1.2 Data preprocessing

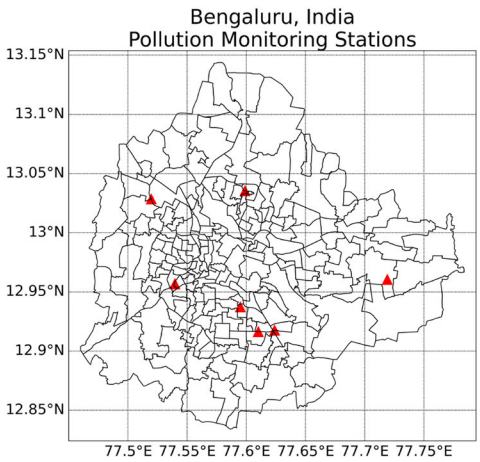
The $PM_{2.5}$ data extracted from Kaggle [31] over the time-period of August 2018 to February 2020 contained several missing values at random, which were imputed over a window length $m$ by taking the mean of the pollution values $v_{t-j}$ where $1 \leq j \leq m$. In this study, as weekly periodic patterns were observed in the pollution data for all the cities, an appropriate window length $m = 7$ was fixed, and imputations were performed based on a 7-day window.


Kolkata, India
Pollution Monitoring Stations
22.6°N
22.55°N
22.5°N
22.45°N
88.25°E
88.3°E
88.35°E
88.4°E
88.45°E

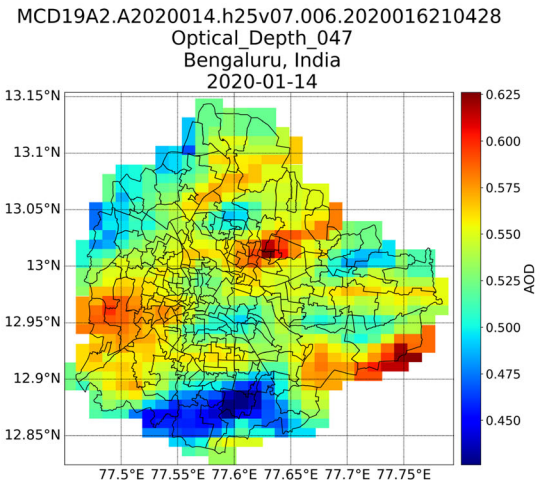
(a)


MCD19A2.A2020016.h26v06.006.2020018044517
Optical_Depth_047
Kolkata, India
2020-01-16
22.6°N
22.55°N
22.5°N
22.45°N
88.25°E
88.3°E
88.35°E
88.4°E
88.45°E
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
AOD

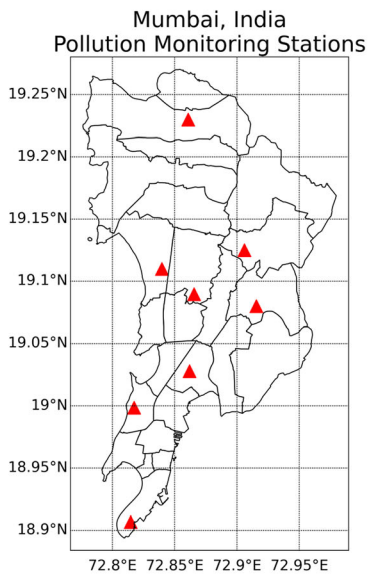
(b)


Bengaluru, India
Pollution Monitoring Stations
13.15°N
13.1°N
13.05°N
13°N
12.95°N
12.9°N
12.85°N
77.5°E
77.55°E
77.6°E
77.65°E
77.7°E
77.75°E

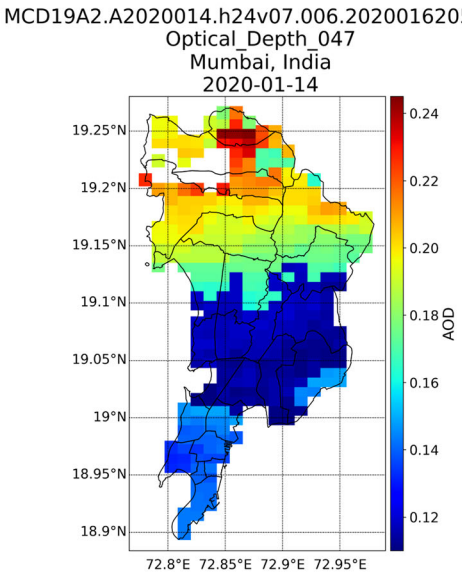
(c)


MCD19A2.A2020014.h25v07.006.2020016210428
Optical_Depth_047
Bengaluru, India
2020-01-14
13.15°N
13.1°N
13.05°N
13°N
12.95°N
12.9°N
12.85°N
77.5°E
77.55°E
77.6°E
77.65°E
77.7°E
77.75°E
0.625
0.600
0.575
0.550
0.525
0.500
0.475
0.450
AOD

(d)


Mumbai, India
Pollution Monitoring Stations
19.25°N
19.2°N
19.15°N
19.1°N
19.05°N
19°N
18.95°N
18.9°N
72.8°E
72.85°E
72.9°E
72.95°E

(e)


MCD19A2.A2020014.h24v07.006.2020016205917
Optical_Depth_047
Mumbai, India
2020-01-14
19.25°N
19.2°N
19.15°N
19.1°N
19.05°N
19°N
18.95°N
18.9°N
72.8°E
72.85°E
72.9°E
72.95°E
0.24
0.22
0.20
0.18
0.16
0.14
0.12
AOD

(f)

◀**Fig. 1** Maps of Kolkata, Bengaluru and Mumbai depicting (**a**, **c**, **e**) pollution monitoring stations and (**b**, **d**, **f**) sample AOD values extracted from MODIS, respectively

Since stations which had missing values greater than 10% were eliminated in a pre-processing step, missing values in the remaining stations were mostly distributed randomly with no occurrences of missing in blocks greater than 7 day periods. As the missing at random instances was statistically insignificant, they did not affect the overall trends in the data during the time-periods inspected for different stations.

To aid in model training, the time-series data were converted into a form that is supportive of supervised modelling approaches. To achieve this, a window length $l$ is chosen. The time-series data are then shifted by one period in the backward direction from the last $l$ times with recording of the contents each time the shift is done. A column wise concatenation operation results in tuples of the form $\{x_{t-l}, x_{t-(l-1)}, ..., x_t\}$ from which $x_t$ can be used as the ground-truth.

In the first stage of AOD preprocessing, the AOD data corresponding to different orbital overpasses were averaged and scaled resulting in a two-dimensional 1200 × 1200 grid taken for further processing. To get the AOD value for a specific monitoring station $S_i$ having the latitude and longitude values as ($\phi_i$, $\lambda_i$) in radians, the Haversine distance $d_{i,j}$ (using Eqs. 1 and 2) was first calculated with respect to every ($\phi_j$, $\lambda_j$) pair present in the 1200 × 1200 grid.

$$h_{ij} = \sqrt{\sin^2\left(\frac{\phi_j - \phi_i}{2}\right) + \cos(\phi_j)\cos(\phi_i)\sin^2\left(\frac{\lambda_i - \lambda_j}{2}\right)} \tag{1}$$

$$d_{ij} = 2r\arcsin\left(\sqrt{h_{ij}}\right) \tag{2}$$

The pair having the minimum distance was taken and the grid indices were found out. Based on the grid indices and also considering the possibility of NaN value being present, a 3 × 3 sub-grid $G$ was taken and the mean of all the values was taken as the AOD value $\tau_{i_t}$ at time instant $t$ for station $S_i$. The formula used to average the sub-grid value for grid indices $(x_i, y_i)$ is presented in Eq. 3 having $\eta$ non-NaN values.

$$\tau_{i_t} = \frac{\sum_{k=x_i-1}^{x_i+1}\left(\sum_{l=y_i-1}^{y_i+1} G_{k,l}\right)}{\eta} \tag{3}$$

#### 4.1.3 Model building

As shown in Fig. 2, the hybrid method $\Psi$ can be denoted as a tuple of 4 models corresponding to different stages in the following form presented in Eq. 4.

$$\Psi = \langle \Gamma, \Theta, \{\Phi_i\}_{i=1}^n, \Pi \rangle \tag{4}$$

where $\Gamma$, $\Theta$ and $\Pi$ are the models for temporal prediction, $PM_{2.5}$ to AOD and AOD to $PM_{2.5}$ conversion, respectively. The set $\{\Phi_i\}_{i=1}^n$ refers to the set of spatial models corresponding to each station $S_i$ for spatial interpolation purposes. To reduce the number of combinations, only the best models for $\Theta$, $\{\Phi_i\}_{i=1}^n$ and $\Pi$ were taken, empirically determined from the results of the experiments performed.

In the first stage involving temporal modelling, a wide variety of models from various types of approaches had been used. From the category of statistical time-series modelling approaches, Auto-Regressive (AR) [35], Auto-Regressive Integrated Moving Average (ARIMA) [35] and Holt-Winters [36] along with deep learning based time-series modelling techniques such as Multi-Layer Perceptron (MLP) [37], Stacked [38], Bi-LSTM [38] and Auto-Encoder [38] coupled with regression techniques such as Linear, Polynomial, Random Forest [39], Support Vector [40] and Decision Tree [41] were used to predict the pollution values over the next $t'$ timesteps using one step-ahead predictions (OSAP) method.

After successfully prediction, in the $PM_{2.5}$ to AOD conversion stage, the forecasted $PM_{2.5}$ values were converted into AOD using a regressor $\Theta$ which could be anyone belonging to Polynomial, Linear, Decision Tree, Random Forest or SVR Regressors. In our experiments, performance of multiple models was investigated and the output from the best performing model for each city was carried forward towards the next stage.

Following $PM_{2.5}$ to AOD conversion, in the spatial interpolation stage, respective spatial models $\Phi_i$ were produced for each station $S_i$. Each spatial model $\Phi_i$ was then used to produce corresponding AOD predictions for each unknown station $S_j$. The intermediate predictions $\Upsilon_i$ produced by each spatial model $\Phi_i$ were weighted based on the inverse squared of the distance $d_{i,j}$ between known $S_i$ and unknown $S_j$ and averaged over the total sum of the weights using Eqs. 5 and 6.

$$W_i = \frac{1}{d_{i,j}^2} \tag{5}$$

$$\Lambda_j = \frac{\sum_{i=1}^{n}(W_i \times \Upsilon_i)}{\sum_{i=1}^{n} W_i} \tag{6}$$

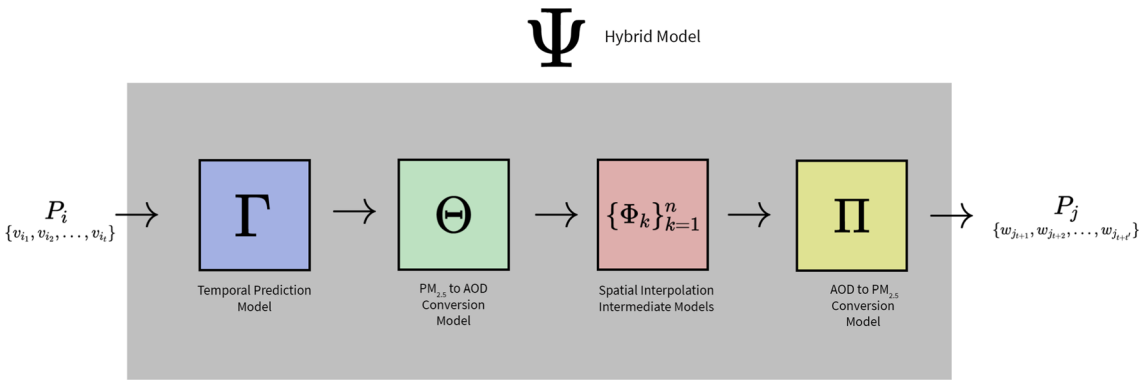



**Fig. 2** Visual depiction of the hybrid model $\Psi$ proposed consisting of temporal prediction model $\Gamma$, $PM_{2.5}$ to AOD regressor $\Theta$, spatial intermediate models $\{\Phi_i\}_{i=1}^{n}$ and AOD to $PM_{2.5}$ regressor $\Pi$. $P_i$ is the input pollution data for the known station $S_i$ while $P_j$ represents the predicted pollution values by hybrid model $\Psi$ for the unknown station $S_j$

where $W_i$ is the weight associated with each station $S_i$ and $\Lambda_j$ is the resulting predictions produced.

In the last stage, the final predictions for $S_j$ were produced by converting the spatially interpolated AOD values $\Lambda_j$ into their corresponding $PM_{2.5}$ values using regressor $\Pi$ constructed similarly to $\Theta$ in the second stage.

In the hybrid approach proposed, it must be noted that spatially interpolation is only conducted over other unmeasured data points in the same city where the temporal forecasting in the first stage has taken place. From extensive experimentations before, it was observed that using spatial models developed for one city to interpolate data over a different city can lead to erroneous results due to several extraneous factors which might become significant due to location and socio-economic differences. Hence, the spatial interpolation third stage should be ideally restricted to a 50 km radius from the geographic center of a city [42] [43].

As models $\Theta$, $\Pi$ and $\Phi_i$ are singular, each hybrid-approach would be recognized by the temporal model $\Gamma$ from now on, in this paper.

## 4.2 Models

To model continuous data, three different types of approaches had been used. In statistical approaches, two variations of AR model (vanilla AR, ARIMA) have been used along with a Triple Exponential Smoothening Technique (Holt-Winters) were used. AR model can described as a linear regression model that takes input, past observations and uses them to produce a prediction for the future time slots. Coefficients of the AR model can be solved by ordinary least squares (OLS) [44] method or through using Yule-Walker [45] equations. An extension of the AR model happens to be the ARIMA model which contains three parts : AR, I and MA. The AR component stands for the auto-regressive terms present in the model. The I component specifies the differencing step required to convert non-stationary time-series to stationary. The MA component resembles the linear combination of errors in past time-steps. Different from traditional AR-based approaches, the Holt-Winters model is an approach that makes use 4 equations, namely a forecast equation and three smoothing equations, and can be represented in additive as well as multiplicative manners. In the current study since seasonal fluctuations did not change the level of the time-series considered, the additive method was considered.

Under shallow machine learning, different types of regression approaches were used. Linear regression is an approach to modelling the association between independent scalar variables $x_i$ and dependent variable $y$, under the important assumption that the relationship between the two different types of variables is linear. Similar to linear regression, polynomial regression uses $\beta_p$ and $\varepsilon$ terms but by treating different powers of the independent variable $x$ as independent terms. In case of SVR, a maximal margin hyperplane is determined through an algorithm [46] which focuses on having errors within a specific margin rather than minimizing the error itself in contrast to other regression models. In feature space transformation, a variety of kernels are used. For this study, through extensive experimentations, the popular radial basis function kernel was found to yield the best results.

Decision Trees create a predictive model which maps observations about items to conclusions. Although decision tree approaches are mostly used for classification problems, by proper data modelling, it can be used to produce meaningful continuous output. To measure the efficiency, various metrics are used of which Variance Reduction [47] is used in case of regression trees to strategically determine the proper splits to minimize variance. Random Forest is an ensemble method built on top of decision trees, where multiple trees are built by frequently resampling training

samples with a random subset of the features and uses averaging to control over-fitting and enhance predictive accuracy.

In the case of deep learning, variations of LSTM along with a MLP method were used. Long short-term memory (LSTM) [48] [49] networks are considered to be modified version of recurrent neural networks (RNN). Unlike RNNs, they do not suffer from the vanishing and exploding gradient problems. The auto-encoder version of LSTM learns the encoding for a dataset through a reduction and reconstruction manner. After completing input reduction, reconstruction is performed where the network tries to generate output similar to the input as possible. A bi-directional LSTM is a form of recurrent neural networks [50] where 2 hidden layers are linked to the paired output having opposite directions. This benefits the output layer to leverage from both the future and past time steps simultaneously. Unlike LSTMs, MLPs belong to the class of feedforward artificial neural networks. It includes at least 3 major layers namely, input layer, hidden layer and an output layer.

## 4.3 Data modelling

### 4.3.1 Time series modelling

The approach for modelling time-series data as depicted in Fig. 3 required fitting the data on a training set and carrying out OSAP on the test/validation set as applicable. The temporal model $\Gamma$ was re-trained in each run on new data. The new data were created by appending the predicted data in the $i-1$ th run to the existing data $P_{i_{in}}$ provided as initial input. For each run, the resulting prediction $w_{i_k}\,\forall k \in \{t+1, t+2, ..., t+t'\}$ was appended to the result set $P_{i_{out}}$ produced as the required output. The predicted values produced serves the basis for the conversion to equivalent AOD and spatial interpolation process in later stages.

### 4.3.2 Inter-conversion between $PM_{2.5}$ and AOD

As shown in Fig. 3, to create both regressor $\Theta$ for the conversion of $PM_{2.5}$ into AOD and regressor $\Pi$ to convert AOD back into $PM_{2.5}$, a similar approach was used. The independent variable as applicable was labeled $X$ and the dependent variable was termed $y$. Based on the data extracted, the AOD values and the $PM_{2.5}$ for each time instant $t$ was found out and put under the respective variables i.e. $X$ and $y$ . $\Theta$ and $\Pi$ were then fitted on $X$ based on $y$ acting as the ground truth.

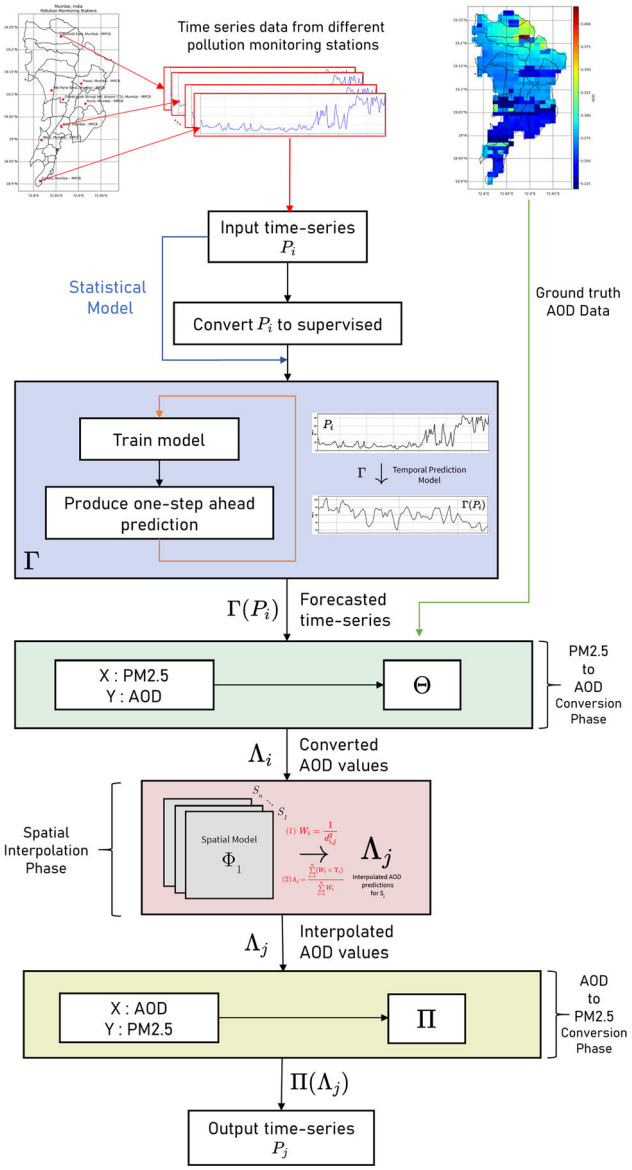



**Fig. 3** Flow diagram to demonstrate the modelling of data and the resulting output

### 4.3.3 Spatial interpolation modelling

In this stage, the raw data were first converted into tuples which were then used to create the independent $X$ and dependent $y$ sets. For modelling spatial data, a station $S_f$ was first fixed $\forall f \in \{1, 2, 3, ..., n\}$. The AOD value at each timestep $i$ was used as the $y$ value of the row, while the tuple containing the AOD value of the fixed station $S_f$ and difference in latitude and longitude of the fixed station $S_f$ w.r.t $S_j\,\forall j \in \{1, 2, 3, ..., i, ..., n\}$ was taken as the $X$ value. The $X$ and $y$ sets so created based on the fixed station $S_f$ were then fed into corresponding intermediate model $\Phi_f$ for training. After training the predictions made $\Upsilon_f$ were processed as per Eqs. 5 and 6 to produce $\Lambda_j$ for the

unknown station $S_j$ which were then converted into $P_j$ using regressor $\Pi$.

## 5 Evaluation

### 5.1 Experimental setup

All model computations had been performed on a Linux-based setup consisting of a 16 GB 3000 MHz DDR4 RAM set coupled with a 6 core 3.6 GHz CPU. To speed up matrix calculations and increase parallel processing, a RTX 3080 GPU was used.

The language of choice for the code used in this study had been Python [51] due to the rich ecosystem consisting of several highly efficient libraries such as TensorFlow [52], StatsModels [53] and Sci-kit Learn [54]. The presence of such libraries was effective in reducing overall complexity of the code without any trade off with regards to performance.

The time-series data used for this study were separated into three groups, namely train, validation and test sets. Train sets involved time-series pollution data from June 2019 to November 2019. Validation set required to test the hyper-parameters spanned over the month of December 2019 while the test set which was used to report the model performance ranged for a period of two months from January 2020 to February 2020. For the purpose of inter-conversion, AOD and $PM_{2.5}$ values ranging over the period June to November 2019 were taken into consideration. The spatial modelling was done on AOD values ranging over the interval one year before the test set period, i.e., over January 2019 to February 2019. The choice of this interval was due to the presence of spatial patterns in Indian cities that are seasonal in nature. The stations were also divided into two sets labelled known and unknown sets chosen out of random in a ratio of 2:1.

Finally for evaluation purposes, the performances of different models were assessed based on the following four metrics namely, Root Mean Squared Error (RMSE) [55], Mean Average Error (MAE) [55], Mean Squared Log Error (MSLE) [56] and Median Absolute Error (MedAE) [57].

### 5.2 Intermediate results

#### 5.2.1 Temporal modelling

Multiple models from different kinds of approaches were used in performing one-step ahead predictions of the input $PM_{2.5}$ time-series values resulting in non-identical performance depending on various cities.

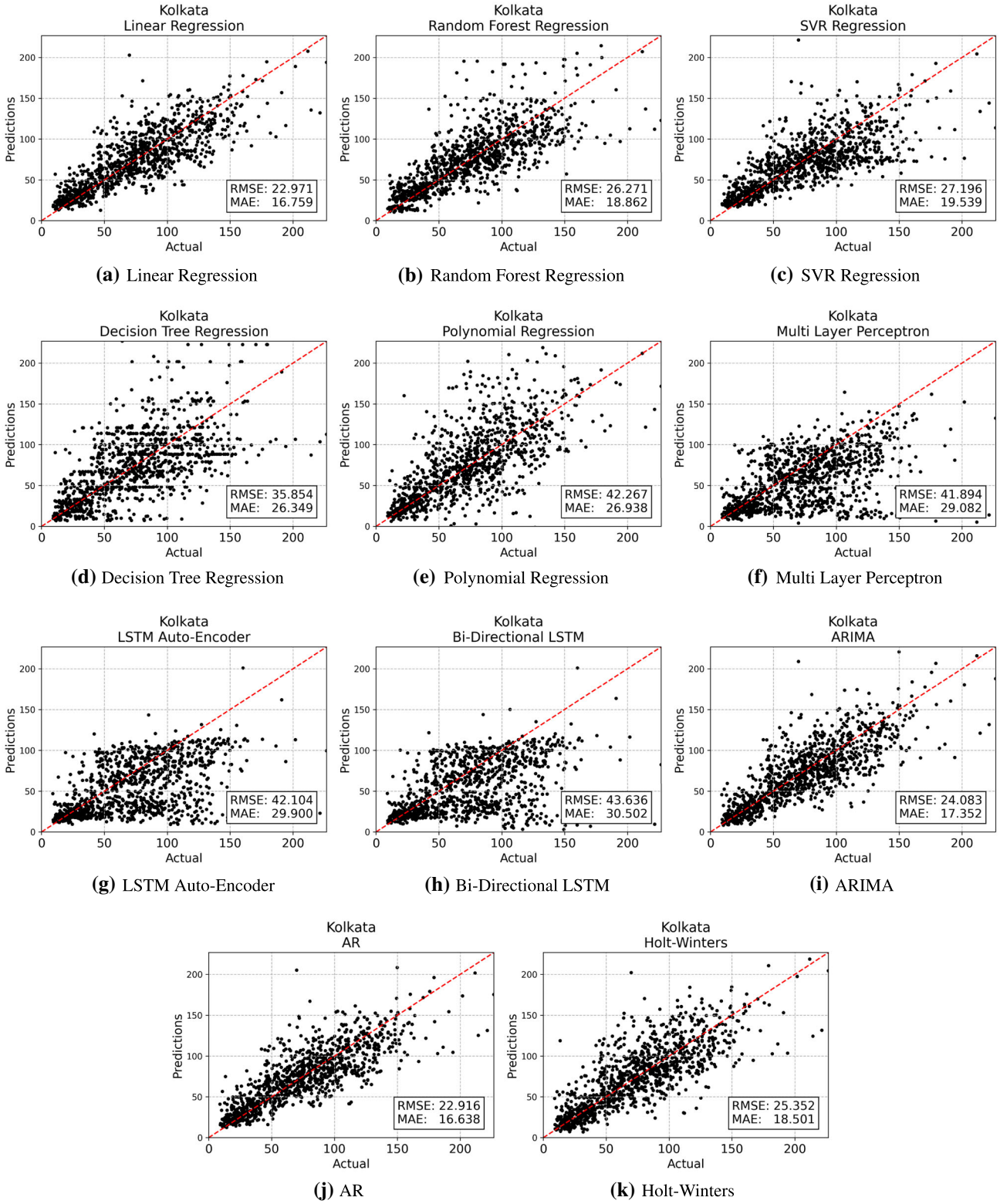
In case of Kolkata, from the correlation plots in Fig. 4, it can be seen that AR (Fig. 4j) performed the best with an RMSE and MAE of 22.916 and 16.638, respectively, followed closely by Linear Regression (Fig. 4a) with RMSE and MAE of 22.971 and 16.759, respectively. From the difference in RMSEs of different models, it is clearly evident that statistical-based approaches outperformed all deep learning and machine learning-based approaches in terms of one step ahead $PM_{2.5}$ forecasting for the city of Kolkata. For LSTM-based approaches such as LSTM Auto-Encoder (Fig. 4g) and Bi-Directional LSTM (Fig. 4h), an evidence of saturation after the point of 100 can be observed, while the MLP-based approach (Fig. 4f) does not show such kind of behavior. This saturation can serve as an indicator of high irregularity that LSTM-based models might not be able to properly track.

Unlike Kolkata, in case of Bengaluru, the best performing model in predicting $PM_{2.5}$ values turned out to be SVR Regression (Fig. 5c) with an RMSE and MAE of 27.403 and 10.316, respectively. Another contrasting observation that can be seen from the correlation plots in Fig. 5 is the increased level of performance in case of deep learning models which supposedly saturated due to irregularities in case of Kolkata. The MLP-based approach (Fig. 5f), fared the second best with an RMSE and MAE of 27.498 and 12.494, respectively, while the different versions of LSTM such as LSTM Auto-Encoder (Fig. 5g) and Bi-Directional LSTM (Fig. 5h) also performed significantly better. This increase in performance can be attributed to low irregularities present in Bengaluru $PM_{2.5}$ values compared to other cities [58].

For the city of Mumbai, it can be observed from the correlation plots in Fig. 6, if efficiency in terms of error is to be evaluated, just like Kolkata, the AR model gave the best MAE and RMSE with values of 15.075 and 19.625, respectively, followed by Linear Regression (Fig. 6a) and Random Forest(Fig. 6b). Statistical Models such as AR (Fig. 6j), ARIMA (Fig. 6i) and Holt-Winters (Fig. 6k) fared much better than their deep learning and machine learning counter parts. Within the category of machine learning models, Polynomial Regression (Fig. 6e) and Decision Tree Regression (Fig. 6d) showed the worst prediction performance performing even worse than all the deep learning models. In deep learning, the MLP approach (Fig. 6f) fared better than LSTM-based approaches such as LSTM Auto-Encoder (Fig. 6g) and Bi-Directional LSTM (Fig. 6h).

#### 5.2.2 $PM_{2.5}$ to AOD and AOD to $PM_{2.5}$ conversion

Although the hierarchy in model performance observed in case of Mumbai and Bengaluru is found to be similar, such was not in the case of Kolkata. For conversion from $PM_{2.5}$



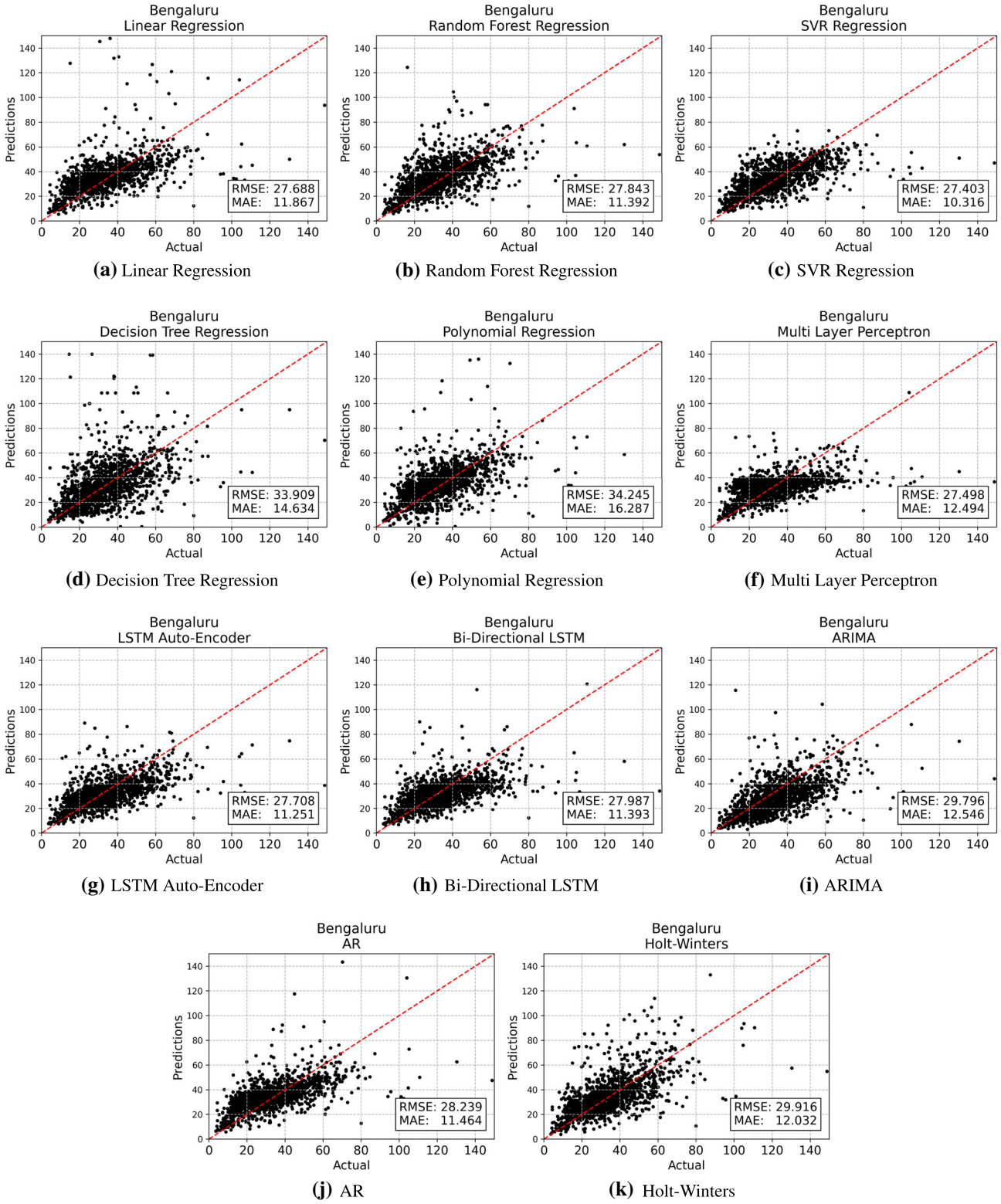

**(a)** Linear Regression **(b)** Random Forest Regression **(c)** SVR Regression

**(d)** Decision Tree Regression **(e)** Polynomial Regression **(f)** Multi Layer Perceptron

**(g)** LSTM Auto-Encoder **(h)** Bi-Directional LSTM **(i)** ARIMA

**(j)** AR **(k)** Holt-Winters

**Fig. 4** Actual vs predicted correlation plots for different temporal models belonging to (**a**–**e**) machine learning, (**f**–**i**) deep learning and (**j**–**l**) statistical approaches for the city of Kolkata




**(a)** Linear Regression **(b)** Random Forest Regression **(c)** SVR Regression

**(d)** Decision Tree Regression **(e)** Polynomial Regression **(f)** Multi Layer Perceptron

**(g)** LSTM Auto-Encoder **(h)** Bi-Directional LSTM **(i)** ARIMA

**(j)** AR **(k)** Holt-Winters

**Fig. 5** Actual vs predicted correlation plots for different temporal models belonging to (**a**–**e**) machine learning, (**f**–**i**) deep learning and (**j**–**l**) statistical approaches for the city of Bengaluru

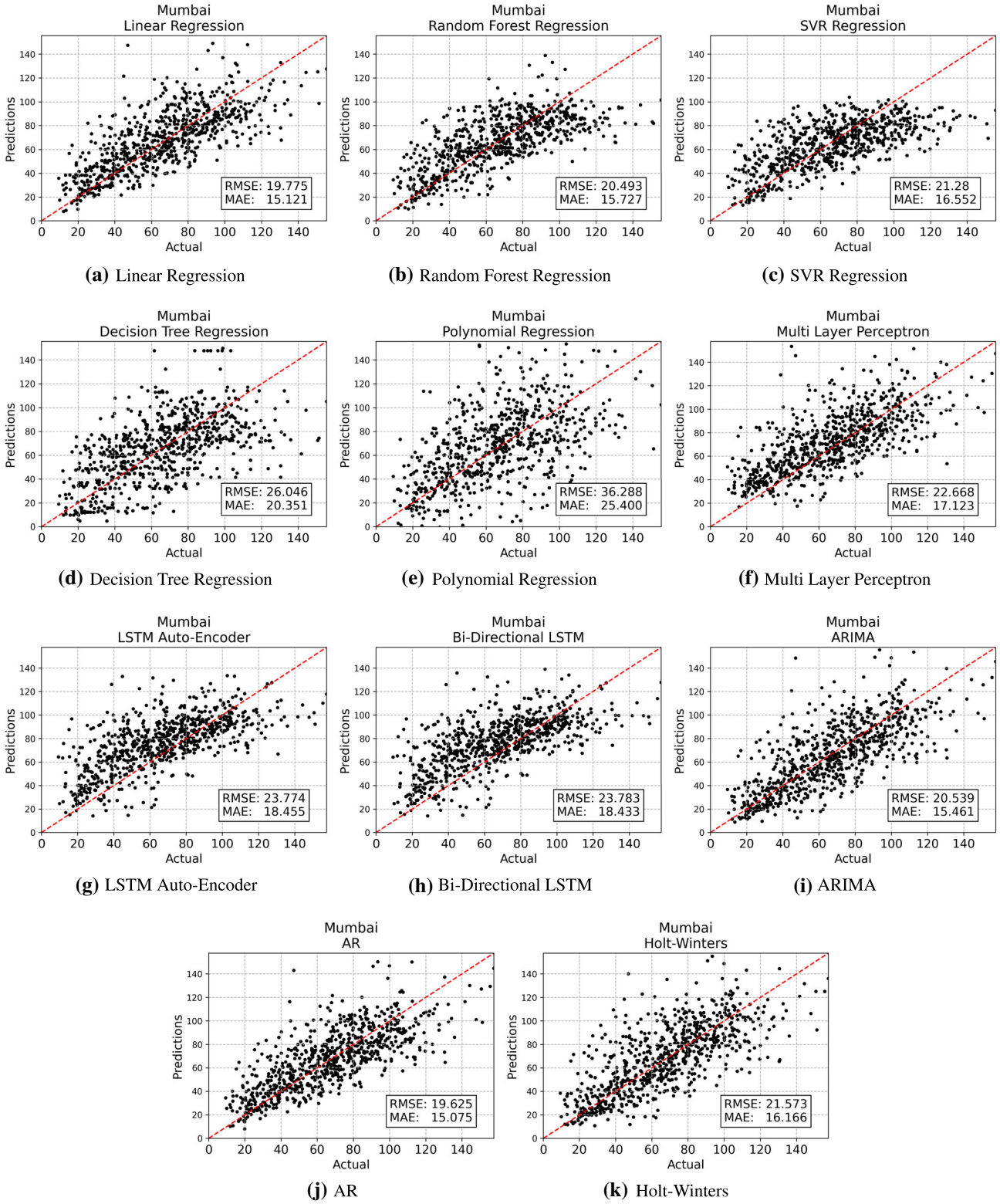



**Fig. 6** Actual vs predicted correlation plots for different temporal models belonging to (**a**-**e**) machine learning, (**f**-**i**) deep learning and (**j**-**l**) statistical approaches for the city of Mumbai

**Table 2** Kolkata : Performance metrics during the Interconversion phase between $PM_{2.5}$ and AOD using various machine learning models

(a) $PM_{2.5}$ to AOD conversion

| Conversion Model ($\Theta$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 0.070 | 0.086 | 0.006 | 0.066 |
| Linear Regression | 0.070 | 0.086 | 0.006 | 0.063 |
| Decision Tree Regression | 0.125 | 0.183 | 0.022 | 0.077 |
| Random Forest Regression | 0.106 | 0.146 | 0.015 | 0.072 |
| SVR Regression | **0.061** | **0.087** | **0.006** | **0.045** |

(b) AOD to $PM_{2.5}$ conversion

| Conversion Model ($\Pi$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 30.798 | 38.919 | 0.128 | 21.469 |
| Linear Regression | **30.377** | **38.612** | **0.125** | **20.714** |
| Decision Tree Regression | 34.529 | 43.097 | 0.173 | 27.030 |
| Random Forest Regression | 31.492 | 39.253 | 0.147 | 25.598 |
| SVR Regression | 32.286 | 40.776 | 0.145 | 23.453 |

**Table 3** Bengaluru : Performance metrics during the Interconversion phase between $PM_{2.5}$ and AOD using various machine learning models

(a) $PM_{2.5}$ to AOD conversion

| Conversion Model ($\Theta$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 0.102 | 0.151 | 0.008 | 0.072 |
| Linear Regression | 0.101 | 0.148 | 0.008 | 0.075 |
| Decision Tree Regression | 0.132 | 0.178 | 0.012 | 0.088 |
| Random Forest Regression | 0.118 | 0.163 | 0.010 | 0.091 |
| SVR Regression | **0.088** | **0.142** | **0.007** | **0.054** |

(b) AOD to $PM_{2.5}$ conversion

| Conversion Model ($\Pi$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 14.227 | **20.020** | 0.260 | 10.729 |
| Linear Regression | 14.229 | 19.398 | 0.259 | 12.114 |
| Decision Tree Regression | 17.380 | 21.285 | 0.367 | 15.740 |
| Random Forest Regression | 16.192 | 20.539 | 0.318 | 14.378 |
| SVR Regression | **13.736** | 20.209 | **0.248** | **9.318** |

to AOD corresponding to model $\Theta$ in our four-staged approach, SVR Regression is found to outperform all other models under the machine learning category consistently in all the cities as can be evident from Tables 2a, 3a and 4a in all the performance metrics used for evaluation. However in case of conversion from AOD to $PM_{2.5}$ as can be seen

**Table 4** Mumbai : Performance metrics during the Interconversion phase between $PM_{2.5}$ and AOD using various machine learning models

(a) $PM_{2.5}$ to AOD conversion

| Conversion Model ($\Theta$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 0.095 | 0.114 | 0.008 | 0.086 |
| Linear Regression | 0.098 | 0.117 | 0.008 | 0.092 |
| Decision Tree Regression | 0.120 | 0.157 | 0.015 | 0.092 |
| Random Forest Regression | 0.097 | 0.129 | 0.010 | 0.075 |
| SVR Regression | **0.089** | **0.113** | **0.008** | **0.073** |

(b) AOD to $PM_{2.5}$ conversion

| Conversion Model ($\Pi$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 22.917 | **28.916** | 0.329 | 18.817 |
| Linear Regression | 24.045 | 29.670 | 0.346 | 22.207 |
| Decision Tree Regression | 24.742 | 32.865 | 0.476 | 19.335 |
| Random Forest Regression | 22.320 | 29.925 | 0.385 | 20.338 |
| SVR Regression | **22.206** | 29.463 | **0.327** | **18.386** |

from Tables 2b, 3b and 4b, Linear Regression is found to outperform in Kolkata, while SVR Regression outperforms others in Bengaluru and Mumbai except in terms of RMSE where Polynomial Regression is found to have a slight advantage.

In terms of error while conversion from $PM_{2.5}$ to AOD denoted by $\Pi$ in our proposed four-staged approach, minimum error in case of MAE and RMSE can be seen in case of Bengaluru with values of 13.736 and 20.020, respectively. Kolkata suffers the most with the best performing model having an RMSE and MAE of 38.612 and 30.377, respectively. As mentioned before, to reduce complexity in terms of different combinations, the best performing model corresponding to each city is taken as the conversion model denoted by $\Theta$ and $\Pi$, respectively.

### 5.2.3 Spatial interpolation

Since reporting the performance for all the machine learning models for every fixed station $S_f$ belonging to every city, the mean of the performance metrics for all the models over all the stations is reported in Tables 5a, 5b and 5c where it can be seen that spatial interpolation of AOD follow a similar pattern in terms of performance hierarchy in contrast to temporal modelling, where each city almost had a unique pattern. As can be seen from Tables 5a, 5b and 5c, Random Forest Regression is observed to outperform all other machine learning-based models in all the performance metrics except in MedAE where Decision

**Table 5** Mean spatial interpolation performance of $\{\Phi_i\}_{i=1}^{n}$ models

(a) Kolkata

| Model ($\Phi$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 0.059 | 0.091 | 0.004 | 0.042 |
| Linear Regression | 0.053 | 0.077 | 0.003 | 0.036 |
| Decision Tree Regression | 0.041 | 0.071 | 0.002 | **0.017** |
| Random Forest Regression | **0.037** | **0.062** | **0.002** | 0.019 |
| SVR Regression | 0.046 | 0.074 | 0.002 | 0.030 |

(b) Bengaluru

| Model ($\Phi$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 0.063 | 0.089 | 0.002 | 0.045 |
| Linear Regression | 0.072 | 0.102 | 0.002 | 0.052 |
| Decision Tree Regression | 0.026 | 0.051 | 0.001 | 0.014 |
| Random Forest Regression | **0.022** | **0.041** | **0.000** | **0.011** |
| SVR Regression | 0.059 | 0.086 | 0.002 | 0.040 |

(c) Mumbai

| Model ($\Phi$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|
| Polynomial Regression | 0.055 | 0.077 | 0.002 | 0.036 |
| Linear Regression | 0.056 | 0.079 | 0.002 | 0.037 |
| Decision Tree Regression | 0.030 | 0.063 | 0.001 | **0.006** |
| Random Forest Regression | **0.028** | **0.050** | **0.001** | 0.009 |
| SVR Regression | 0.052 | 0.077 | 0.002 | 0.031 |

Tree Regression had a slight edge in Kolkata and Mumbai. The short range of error values for the three cities suggests that the surface on which spatial interpolation is plain and does not possess significant undulations to introduce irregularities.

The best performing model producing the minimum mean MAE and RMSE performance is denoted by $\Phi$ and is chosen for spatial interpolation for every fixed station $S_f$, to produce the required set of $\{\Phi_i\}_{i=1}^{n}$ models for our proposed four-staged approach.

## 5.3 Final results

### 5.3.1 Kolkata

As can be seen in Table 6a, Random Forest Regression when used as the temporal model $\Gamma$ shows the best performance in terms of MAE, RMSE and MedAE with values 27.993, 37.633 and 18.458, respectively. In terms of MSLE, ARIMA model outperformed others although with a small margin. Under the deep learning category, methods using Bi-Directional and LSTM Auto-Encoder show similar performance to Random Forest Regression with little difference in MAE, RMSE and MedAE values. In case of statistical-based approaches the AR- and Holt-Winters-based models performed slightly better than ARIMA in terms of both RMSE and MAE with values 27.998 and 37.645, respectively, but only to perform a little worse in MSLE. Even though there are little differences, it can be observed from the relative improvement graphs in Fig. 7 that almost all methods at the end perform very similar to each other with changes not increasing more than 0.06% as shown in Fig. 7b.

### 5.3.2 Bengaluru

In case of Bengaluru, as evident from Table 6b, the MLP-based method shows the lowest error in terms of RMSE and MSLE with a value of 12.290 and 0.188, respectively. Decision Tree Regression-based method performs outperform others in terms of average error with a value of 10.301 while in terms of MedAE, ARIMA is found to outperform all methods with a value of 10.346. Under the statistical category, ARIMA-based method is found to fare the best with an RMSE of 10.307 followed by Holt-Winters and AR-based hybrid methods. Just like in the case of Kolkata, the relative improvements of even the best performing methods with respect to the worst performing ones (Fig. 8) are not significant as the values do not exceed 1% as can be seen in Fig. 8a.

### 5.3.3 Mumbai

From the performance metrics shown in Table 6c, it can be observed that the hybrid technique $\Psi$ using LSTM Auto-Encoder as $\Gamma$ showed the best performance in terms of the RMSE with a value of 25.533. In terms of machine learning-based models, the method with Random Forest Regression outperformed all others with MAE and MSLE values of 20.461 and 0.156, respectively. The AR model-based method performed best among others in the statistical group with an RMSE value of 25.677. On careful examination of performance metrics presented in Table 6c and the relative improvements of different methods are shown in form of bar graphs Fig. 9. , the maximum deviation between the best and the worst performing hybrid method for Mumbai is found to 1.007 and 0.973 in MAE and RMSE values, respectively. Even the maximum relative improvement is not more than 5% as can be seen from the Machine Learning relative improvement graph in Fig. 9a.

**Table 6** Final performance comparison of different hybrid methods recognized on the basis of the temporal model used

(a) Kolkata

[$\Theta$ = SVR Regression, $\Pi$ = Linear Regression and $\Phi$ = Random Forest Regression]

| Approach | Temporal Model ($\Gamma$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|---|
| Machine Learning | SVR Regression | 27.998 | 37.641 | 0.137 | 18.464 |
| | Linear Regression | 28.001 | 37.650 | 0.138 | 18.466 |
| | Random Forest Regression | **27.993** | **37.633** | 0.139 | **18.458** |
| | Polynomial Regression | 27.998 | 37.644 | 0.136 | 18.472 |
| | Decision Tree Regression | 27.994 | 37.639 | 0.139 | 18.458 |
| Deep Learning | Multi-Layer Perceptron | 28.010 | 37.657 | 0.140 | 18.472 |
| | Bi-Directional LSTM | 27.993 | 37.634 | 0.137 | 18.458 |
| | LSTM Auto-Encoder | 27.994 | 37.636 | 0.136 | 18.462 |
| Statistical | AR | 27.998 | 37.645 | 0.140 | 18.464 |
| | Holt-Winters | 27.998 | 37.645 | 0.136 | 18.468 |
| | ARIMA | 28.001 | 37.648 | **0.133** | 18.463 |

(b) Bengaluru

[$\Theta$ = SVR Regression, $\Pi$ = SVR Regression and $\Phi$ = Random Forest Regression]

| Approach | Temporal Model ($\Gamma$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|---|
| Machine Learning | Polynomial Regression | 10.400 | 12.369 | 0.190 | 10.578 |
| | SVR Regression | 10.366 | 12.369 | 0.190 | 10.578 |
| | Linear Regression | 10.319 | 12.333 | 0.189 | 10.703 |
| | Decision Tree Regression | **10.301** | 12.325 | 0.189 | 10.633 |
| | Random Forest Regression | 10.412 | 12.438 | 0.192 | 10.679 |
| Deep Learning | Multi-Layer Perceptron | 10.316 | **12.290** | **0.188** | 10.397 |
| | Bi-Directional LSTM | 10.320 | 12.302 | 0.189 | 10.530 |
| | LSTM Auto-Encoder | 10.335 | 12.329 | 0.189 | 10.475 |
| Statistical | Holt-Winters | 10.309 | 12.333 | 0.189 | 10.462 |
| | ARIMA | 10.307 | 12.299 | 0.188 | **10.346** |
| | AR | 10.369 | 12.381 | 0.190 | 10.594 |

(c) Mumbai

[$\Theta$ = SVR Regression, $\Pi$ = SVR Regression and $\Phi$ = Random Forest Regression]

| Approach | Temporal Model ($\Gamma$) | MAE | RMSE | MSLE | MedAE |
|---|---|---|---|---|---|
| Machine Learning | Polynomial Regression | 21.040 | 26.254 | 0.164 | **17.623** |
| | SVR Regression | 21.092 | 26.029 | 0.161 | 17.823 |
| | Random Forest Regression | **20.461** | 25.590 | **0.156** | 18.468 |
| | Linear Regression | 21.187 | 26.198 | 0.163 | 18.277 |
| | Decision Tree Regression | 21.468 | 26.506 | 0.167 | 18.993 |
| Deep Learning | Multi-Layer Perceptron | 20.977 | 25.955 | 0.160 | 19.018 |
| | Bi-Directional LSTM | 20.563 | 25.574 | 0.158 | 18.006 |
| | LSTM Auto-Encoder | 20.510 | **25.533** | 0.157 | 17.913 |
| Statistical | AR | 20.642 | 25.677 | 0.158 | 18.032 |
| | Holt-Winters | 20.943 | 26.021 | 0.160 | 19.316 |
| | ARIMA | 21.249 | 26.452 | 0.164 | 19.516 |

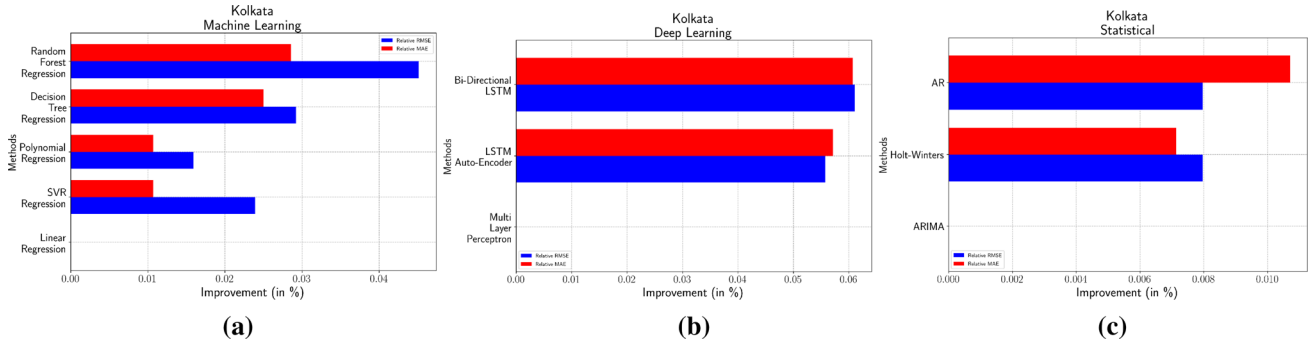



**Fig. 7** Relative improvement graphs of hybrid methods in terms of RMSE and MAE w.r.t. worst performing ones for the city of Kolkata

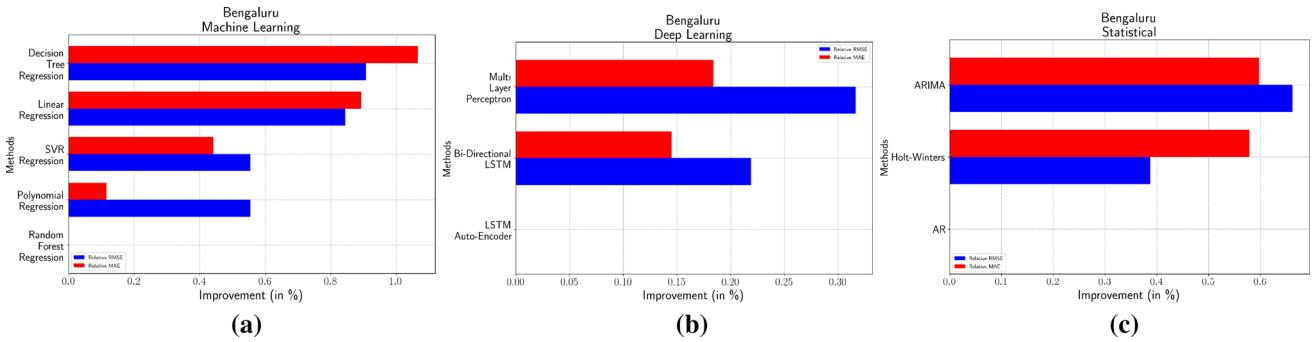



**Fig. 8** Relative improvement graphs of hybrid methods in terms of RMSE and MAE w.r.t. worst performing ones for the city of Bengaluru

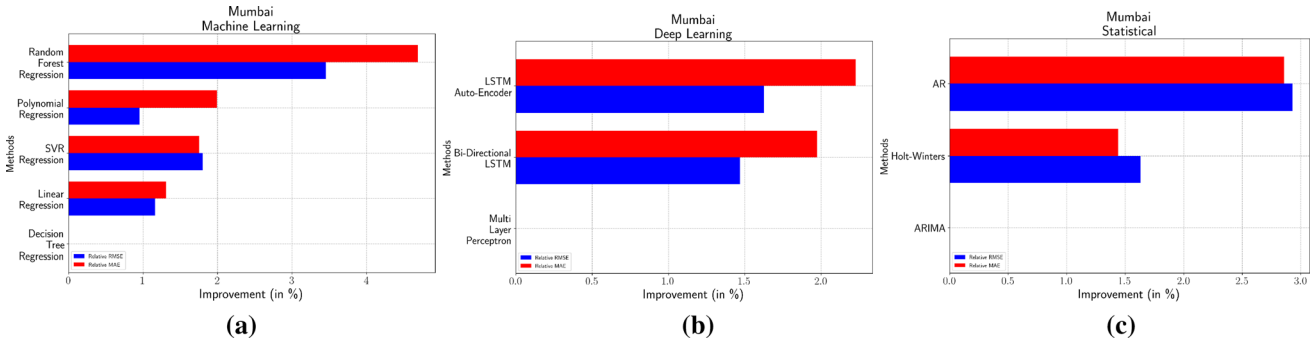



**Fig. 9** Relative improvement graphs of hybrid methods in terms of RMSE and MAE w.r.t. worst performing ones for the city of Mumbai

## 5.4 Discussion

The hybrid approach in this study is proposed as a combination of several baseline models utilizing AOD as an intermediate spatial interpolation medium due to high data availability from space-based satellites setup by NASA. The results in Sect. 5.2 display the performances of different such models in performing the first 3 stages for each city taken into consideration. Having fixed the three city specific models $\Theta, \{\Phi_i\}_{i=1}^{n}$ and $\Pi$, all combinations of $\Gamma$ and the fixed $\langle \Theta, \{\Phi_i\}_{i=1}^{n}, \Pi \rangle$ were tested and performances are shown in Sect. 5.3 as the output of the final stage. As $\Theta, \{\Phi_i\}_{i=1}^{n}$ and $\Pi$ did not have any temporal dependency and can be considered characteristic to the city being investigated, except individual stage specific performances (as shown in Sect. 5.2), the need for testing all combinations was not required other than the one with

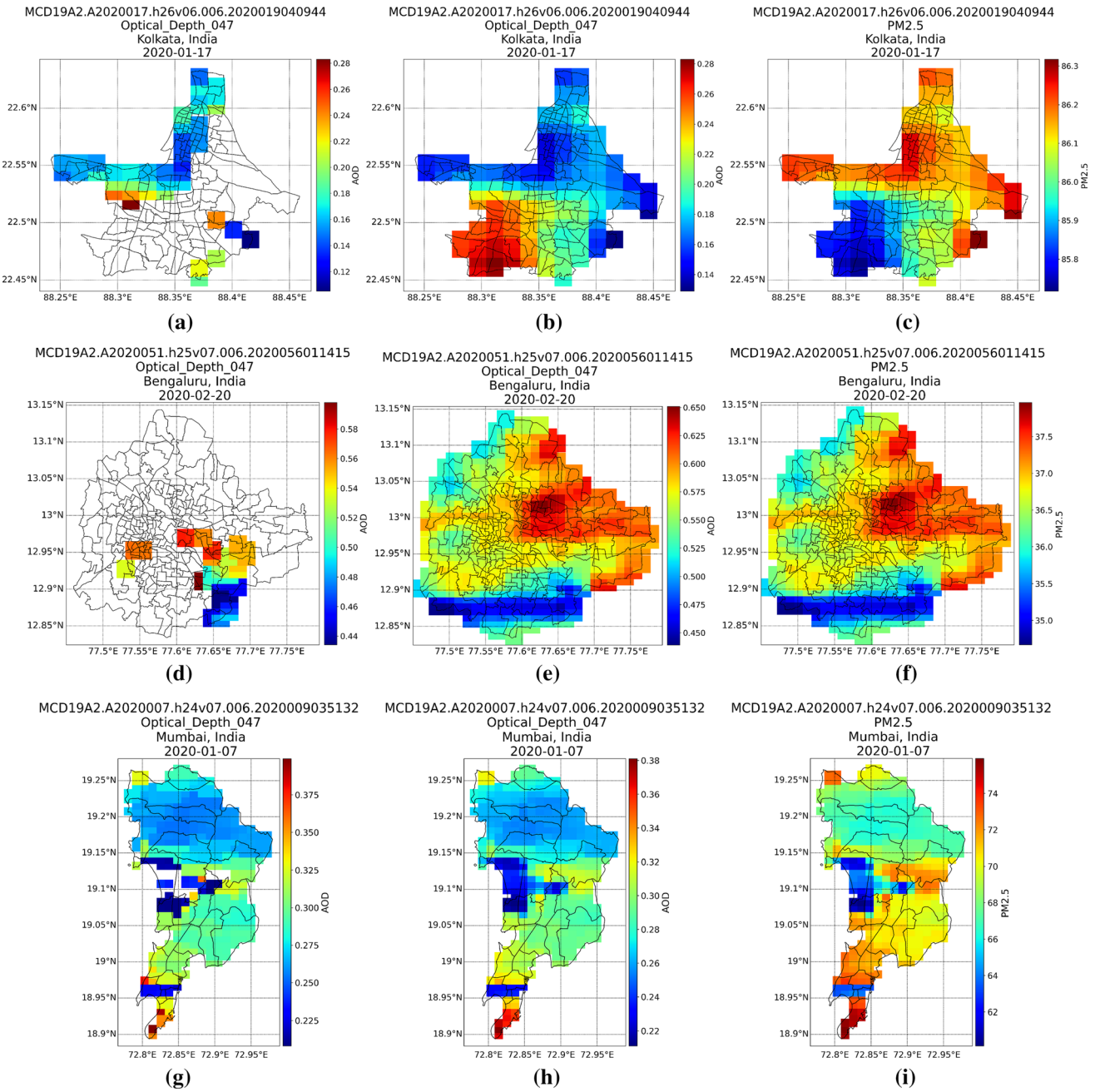



**Fig. 10** Heatmaps for the city of Kolkata, Bengaluru, Mumbai (**a**, **d**, **g**) before interpolation, (**b**, **e**, **h**) after interpolation and (**c**, **f**, **i**) after conversion from AOD to $PM_{2.5}$

various combinations of temporal model $\Gamma$ for a specific city.

Observing the results in sect. 5.2 and 5.3, as the difference in the best performing and the worst performing models does not exceed the 10% mark, the relative changes can be termed insignificant for forecasting related purposes. Also since these values are low compared to the mean MAE and mean RMSE of all the approaches, in terms of implementation if certain models such as LSTM Auto-Encoder becomes computationally expensive, simpler models can also be used with relatively smaller differences in errors.

Another important point that must be noted is that the lack of preservation of temporal model hierarchy such as in case of Mumbai as presented in Fig. 6 and Table 6c can be heavily attributed to the intermediate conversion approaches such as $\Theta$, $\Pi$ and $\{\Phi_i\}_{i=1}^{n}$ involved in the pipeline before producing the final result. Also, the hierarchy in model performance so established depends heavily on the data set used for carrying out the analysis. If a different

dataset is used, the model performance hierarchy can change and may lead to better performance in complex models.

The intermediate process of interpolation of AOD and subsequent conversion to $PM_{2.5}$ data for Kolkata, Bengaluru and Mumbai is visually depicted through heatmaps in Fig. 10. Figs. 10a, 10d and 10g resemble the AOD values originally taken from MODIS, which is interpolated using spatial models $\Phi_i$ as per Eqs. 5 and 6 to produce Figs. 10b, 10e and 10h, respectively. Figs. 10c, 10f and 10i represents the heatmap resulting from the conversion of the interpolated AOD values back into $PM_{2.5}$ using regressor $\Pi$. The efficacy of the interpolation process can be identified from the difference in AOD values at already present sites between the original and the interpolated result. From Fig. 10g and 10h, the change in AOD values can be found to be almost negligible indicated that the interpolation process has been fruitful.

## 6 Conclusion

In this study, a novel class of hybrid spatio-temporal models have been proposed to aid in predicting pollution values of unknown locations taking the help of AOD data for spatial interpolation when historical time-series values of other locations are known. At this stage of the study, only the AOD value along with the corresponding geographic coordinates were taken as features for spatial interpolation and in case of temporal forecasting, only the historical $PM_{2.5}$ values were taken into consideration. Even though complex ML models or DL solutions could have been utilized for spatial interpolation purposes mentioned in the third stage of the approach, the current focus was to devise an effective methodology to produce forecasts that are easy to implement and do not require extensive computation or input data to train. As a possible future work, exogenous variables can be taken into account not only for both spatial interpolation and time-series forecasting purposes, but also in the inter-conversion between $PM_{2.5}$ and AOD.

It is our hope that the hybrid approach proposed in this paper enables the statutory agencies to overcome the problems faced due to data sparsity and the difficulties in establishing air quality monitoring stations at every part of the city, thereby saving millions of dollars in operational and real-estate costs. Even in case of instrument failure, we hope that these networks will come in handy to compensate for the missing pollution readings till the instrument undergoes maintenance and brought back to proper working conditions.

**Acknowledgements** The research work of Asif Iqbal Middya is partially supported by UGC-NET Junior Research Fellowship (UGC-Ref. No.:3684 / (NET-JULY 2018)) provided by the University Grants Commission, Government of India. This research work is also supported by the project entitled "Participatory and Realtime Pollution Monitoring System For Smart City, funded by Higher Education, Science & Technology and Biotechnology, Department of Science & Technology, Government of West Bengal, India".

**Data and Code Availability** The analysis code and data used for this paper upon publication can be found in the following link - https://github.com/nathzi1505/AOD-Hybrid-Paper.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Kennedy D, Bates RR, Watson AY, et al. (1988) Air pollution, the automobile, and public health
2. Paulos E, Anderson K, Townsend A (2004) Ubicomp in the urban frontier. Speech at the Sixth International Conference on Ubiquitous Computing Workshop
3. World Health Organisation (2016) WHO Global Urban Ambient Air Pollution Database. URL https://www.who.int/airpollution/data/cities-2016/en/. Accessed: 2021-06-01
4. The Hindustan Times (2017) Delhi gets 18 more monitoring stations to keep tab on air quality. URL https://tinyurl.com/3mya2zz3. Accessed: 2021-06-01
5. Xing YF, Xu YH, Shi MH, Lian YX (2016) The impact of pm2. 5 on the human respiratory system. J thoracic dis 8(1):E69
6. World Health Organisation (2013) Health effects of Particulate Matter. URL https://www.euro.who.int/__data/assets/pdf_file/0006/189051/Health-effects-of-particulate-matter-final-Eng.pdf. Accessed: 2021-06-01
7. Wang W, Guo Y (2009) Air pollution pm2.5 data analysis in los angeles long beach with seasonal arima model. 2009 Int Conf Energy and Environ Technol 3:7–10
8. Lei F, Dong X, Ma X (2020) Prediction of pm2. 5 concentration considering temporal and spatial features: A case study of fushun, liaoning province. Journal of Intelligent & Fuzzy Systems (Preprint), 1–11
9. Wang M, Sampson PD, Hu J, Kleeman M, Keller JP, Olives C, Szpiro AA, Vedal S, Kaufman JD (2016) Combining land-use regression and chemical transport modeling in a spatiotemporal geostatistical model for ozone and pm2. 5. Environ sci technol 50(10):5111–5118
10. Shao P, Xin J, An J, Kong L, Wang B, Wang J, Wang Y, Wu D (2017) The empirical relationship between pm2. 5 and aod in nanjing of the yangtze river delta. Atmos Pollut Res 8(2):233–243
11. Bui TC, Kim J, Kang T, Lee D, Choi J, Yang I, Jung K, Cha SK (2020) Star: Spatio-temporal prediction of air quality using a multimodal approach
12. He Z, Chow C, Zhang J (2020) Stnn: A spatio-temporal neural network for traffic predictions. IEEE Transactions on Intelligent Transportation Systems pp 1–10
13. Pu Q, Yoo EH (2020) Spatio-temporal modeling of pm2.5 concentrations with missing data problem: a case study in beijing, china. Int J Geogr Inf Sci 34(3):423–447. https://doi.org/10.1080/13658816.2019.1664742

14. Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, Sabath MB, Choirat C, Koutrakis P, Lyapustin A, Wang Y, Mickley LJ, Schwartz J (2019) An ensemble-based model of pm2.5 concentration across the contiguous united states with high spatiotemporal resolution. Environment International **130**:104,909. https://doi.org/10.1016/j.envint.2019.104909. URL https://www.sciencedirect.com/science/article/pii/S0160412019300650
15. Stafoggia M, Bellander T, Bucci S, Davoli M, de Hoogh K, de' Donato F, Gariazzo C, Lyapustin A, Michelozzi P, Renzi M, Scortichini M, Shtein A, Viegi G, Kloog I, Schwartz J (2019) Estimation of daily pm10 and pm2.5 concentrations in italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environment International 124:170–179. https://doi.org/10.1016/j.envint.2019.01.016. URL www.sciencedirect.com/science/article/pii/S0160412018327685
16. Krishna RK, Ghude SD, Kumar R, Beig G, Kulkarni R, Nivdange S, Chate D (2019) Surface PM2.5 estimate using satellite-derived aerosol optical depth over india. Aerosol and Air Quality Res 19(1):25–37. https://doi.org/10.4209/aaqr.2017.12.0568
17. Wu Z, Wang Y, Zhang L (2019) Msstn: Multi-scale spatial temporal network for air pollution prediction. In: 2019 IEEE International Conference on Big Data (Big Data), pp 1547–1556. https://doi.org/10.1109/BigData47090.2019.9005574
18. Lindström J, Szpiro A, Sampson P, Sheppard L, Oron A, Richards M, Larson T (2011) A flexible spatio-temporal model for air pollution with spatio-temporal covariates. ISEE Conference Abstracts 2011. https://doi.org/10.1289/isee.2011.00165
19. Taieb SB, Atiya AF (2016) A bias and variance analysis for multistep-ahead time series forecasting. IEEE Trans Neural Netw Learn Sys 27(1):62–76. https://doi.org/10.1109/TNNLS.2015.2411629
20. Chandra R (2015) Competition and collaboration in cooperative coevolution of elman recurrent neural networks for time-series prediction. IEEE Trans Neural Netw Learn Sys 26(12):3123–3136. https://doi.org/10.1109/TNNLS.2015.2404823
21. Xu M, Yang Y, Han M, Qiu T, Lin H (2019) Spatio-temporal interpolated echo state network for meteorological series prediction. IEEE Trans Neural Netw Learn Sys 30(6):1621–1634
22. Soh P, Chang J, Huang J (2018) Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6:38,186-38,199
23. Zhu JY, Sun C, Li VO (2017) An extended spatio-temporal granger causality model for air quality estimation with heterogeneous urban big data. IEEE Trans on Big Data 3(3):307–319
24. Sahu SK, Gelfand AE, Holland DM (2006) Spatio-temporal modeling of fine particulate matter. Journal of Agricultural, Biological, and Environmental Statistics **11**(1):61–86. URL http://www.jstor.org/stable/27595586
25. Cesario E, Comito C, Talia D (2017) An approach for the discovery and validation of urban mobility patterns. Pervasive and Mobile Computing 42:77–92. https://doi.org/10.1016/j.pmcj.2017.09.006. URL www.sciencedirect.com/science/article/pii/S157411921630390X
26. Comito C (2020) Next: A framework for next-place prediction on location based social networks. Knowledge-Based Systems 204:106,205. https://doi.org/10.1016/j.knosys.2020.106205. URL www.sciencedirect.com/science/article/pii/S095070512030424X
27. Yang Q, Yuan Q, Yue L, Li T, Shen H, Zhang L (2019) The relationships between pm2.5 and aerosol optical depth (aod) in mainland china: About and behind the spatio-temporal variations. Environmental Pollution **248**. https://doi.org/10.1016/j.envpol.2019.02.071
28. Ni X, Cao C, Zhou Y, Cui X, P. Singh R (2018) Spatio-temporal pattern estimation of pm2.5 in beijing-tianjin-hebei region based on modis aod and meteorological data using the back propagation neural network. Atmosphere **9**(3). https://doi.org/10.3390/atmos9030105. URL https://www.mdpi.com/2073-4433/9/3/105
29. Mao X, Shen T, Feng X (2017) Prediction of hourly ground-level pm2. 5 concentrations 3 days in advance using neural networks with satellite data in eastern china. Atmos Pollut Res 8(6):1005–1015
30. Kloog I, Chudnovsky AA, Just AC, Nordio F, Koutrakis P, Coull BA, Lyapustin A, Wang Y, Schwartz J (2014) A new hybrid spatio-temporal model for estimating daily multi-year pm2.5 concentrations across northeastern usa using high resolution aerosol optical depth data. Atmospheric Environment 95:581–590. https://doi.org/10.1016/j.atmosenv.2014.07.014. URL www.sciencedirect.com/science/article/pii/S1352231014005354
31. Rao R. Air quality data in india (2015 - 2020). URL https://www.kaggle.com/rohanrao/air-quality-data-in-india. Accessed: 2021-06-01
32. Ministry of Environment, Forest and Climate Change. Central control room for air quality management. https://cpcb.nic.in/. Accessed: 2021-06-01
33. NASA. MODIS - Moderate Resolution Imaging Spectroradiometer. URL https://terra.nasa.gov/about/terra-instruments/modis. Accessed: 2021-06-01
34. NASA. LAADS DAAC. URL https://ladsweb.modaps.eosdis.nasa.gov/. Accessed: 2021-06-01
35. Connor JT, Martin RD, Atlas LE (1994) Recurrent neural networks and robust time series prediction. IEEE trans neural netw 5(2):240–254
36. Winters PR (1960) Forecasting sales by exponentially weighted moving averages. Manag Sci 6(3):324–342. https://doi.org/10.1287/mnsc.6.3.324
37. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. nature 323(6088):533–536
38. Ma X, Zhang J, Du B, Ding C, Sun L (2018) Parallel architecture of convolutional bi-directional lstm neural networks for network-wide metro ridership prediction. IEEE Trans Intell Transp Sys 20(6):2278–2288
39. Breiman L (2001) Random forests. Mach learn 45(1):5–32
40. Smola AJ, Scholkopf B (2004) A tutorial on support vector regression. Statistics and comput 14(3):199–222
41. Quinlan JR (1987) Simplifying decision trees. Int j man-mach stud 27(3):221–234
42. Griffith DA (2003) Spatial Autocorrelation and Spatial Filtering. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24806-4
43. Exploratory spatial data analysis (esda) and spatial autocorrelation. URL https://cran.r-project.org/web/packages/lctools/vignettes/SpatialAutocorrelation.html. Accessed: 2021-06-01
44. Goldberger AS (1964) Classical linear regression. Econometric theory pp 156–212
45. Walker GT (1931) On periodicity in series of related terms. Proc R Soc London. Series A, Containing Papers of a Math Phys Character 131(818):518–532
46. Chang CC, Lin CJ (2011) Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3). https://doi.org/10.1145/1961189.1961199
47. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press
48. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput. 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
49. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2017) Lstm: A search space odyssey. IEEE Trans Neural Netw Learn Sys 28(10):2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

50. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
51. Van Rossum G, Drake FL Jr (1995) Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands
52. Martín Abadi et al. (2015) Tensorflow:large-scale machine learning on heterogeneous systems
53. Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python
54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
55. Nath P, Saha P, Middya AI, Roy S (2021) Long-term time-series pollution forecast using statistical and deep learning methods. Neural Comput Appl 33(19):12551–12570. https://doi.org/10.1007/s00521-021-05901-2
56. Mean squared logarithmic error (msle): Peltarion platform. URL https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/mean-squared-logarithmic-error-(msle). Accessed: 2021-06-01
57. Sammut C, Webb GI (eds) (2010) Mean Absolute Error, pp 652–652. Springer US, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_525
58. Herald D (2020) How bad is bengaluru air?. URL https://www.deccanherald.com/metrolife/metrolife-your-bond-with-bengaluru/how-bad-is-bengaluru-air-909370.html. Accessed: 2021-06-01