Environmental Science & Technology

Anthropogenic Impacts on the Atmosphere

Estimation of high-resolution PM2.5 over Indo-Gangetic Plain by fusion of satellite data, meteorology, and land use variables

Alaa Mhawish, Tirthankar Banerjee, Meytar Sorek-Hamer, Muhammad Bilal, Alexei Lyapustin, Robert B. Chatfield, and David Broday

Environ. Sci. Technol., **Just Accepted Manuscript** • DOI: 10.1021/acs.est.0c01769 • Publication Date (Web): 03 Jun 2020

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Estimation of high-resolution $PM_{2.5}$ over Indo-Gangetic Plain by fusion of satellite data, meteorology, and land use variables

Alaa Mhawish[1,2,3], Tirthankar Banerjee[3,4]*, Meytar Sorek-Hamer[1,2], Muhammad Bilal[5], Alexei I. Lyapustin[6], Robert Chatfield[2], David M Broday[7]

[1] Universities Space Research Association (USRA), CA, USA
[2] NASA Ames Research Center, Moffett Field, CA, USA
[3]Institute of Environment and Sustainable Development, Banaras Hindu University, Varanasi, India
[4]DST-Mahamana Centre of Excellence in Climate Change Research, Banaras Hindu University, Varanasi, India
[5]School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing, China
[6]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[7]Civil and Environmental Engineering, Technion, Haifa, Israel

*Correspondence to: T. Banerjee (tb.iesd@bhu.ac.in; tirthankaronline@gmail.com)

Key points:

1. High-resolution MAIAC AOD-based $PM_{2.5}$ was estimated over the Indo-Gangetic Plain; fusing satellite data, land-use variables & meteorology.

2. Random forest based AOD-PM model estimates were able to capture and quantify the $PM_{2.5}$ variability at a sub-urban scale.

3. Comparatively high $PM_{2.5}$ concentrations were evident over central and lower IGP, mediated by land-use and local meteorology.

ABSTRACT

Very high spatially resolved satellite-derived ground-level $PM_{2.5}$ concentrations have multiple potential applications especially in air quality modelling, epidemiological and climatological research. Satellite-derived aerosol optical epth (AOD), and columnar water vapor (CWV), meteorological parameters, and land use data were used as variables within a linear mixed effect model (LME) and a random forest (RF) model, to predict daily ground-level concentrations of $PM_{2.5}$ at 1km×1km grid across the Indo-Gangetic Plain (IGP) in South Asia. The RF model exhibited superior performance and higher accuracy than the LME model, with higher cross-validated explained variance (R^2=0.87) and lower relative prediction error (RPE=24.5%). The RF model revealed improved performance metrics for increasing averaging periods, from daily to weekly, monthly, seasonal, and annual means, which supports using it to estimate $PM_{2.5}$ exposure metrics across the IGP at varying temporal scales (i.e. both short and long terms). The RF-based $PM_{2.5}$ estimates show high $PM_{2.5}$ levels over the middle and lower IGP, with the annual mean exceeding 110μg/m^3. Seasonally, winter was the most polluted season while monsoon was the cleanest. Spatially, the middle and lower IGP showed poorer air quality compared to the upper IGP. In winter, the middle and lower IGP experience very poor air quality, with mean $PM_{2.5}$ concentrations >170μg/m^3.

Keywords: Aerosols; Machine Learning; Random Forest; Mixed effect model; MAIAC; IGP.

1. Introduction

Airborne fine particulate matter with an aerodynamic diameter less than 2.5 μm ($PM_{2.5}$) have been associated with many adverse health effects, especially with cardiovascular and respiratory diseases[1]. Numerous epidemiological studies associate exposure to $PM_{2.5}$ with different health outcomes[2–7]. Recently, the World Health Organization estimated around 4.2 million deaths were attributable, globally, to air pollution[8]. However, most of the epidemiological studies have been conducted in major urban areas, where air quality monitoring is denser, rather than in small cities and rural areas. In South Asia, the air quality monitoring stations are sparsely distributed and are found mainly in major cities, such as in Delhi, Mumbai, Dhaka, and in state capitals. In the suburban and rural areas where a major fraction of the population resides, and the $PM_{2.5}$ levels are as high as in urban areas[9–10]; there are only a few air quality monitoring stations, and in some regions, there are none at all. Health risk assessments of $PM_{2.5}$ exposure across highly populated and polluted areas of the Indo-Gangetic Plain (IGP) are severely constrained by the sparse air quality monitoring stations and limited availability of particulate measurement data[11].

Understating the $PM_{2.5}$ spatial and temporal distribution therefore, is essential to improve understanding of its impact on human health and regional climate.

Satellite remote sensing has the capability to provide high spatially resolved aerosol optical depth measurements with daily global coverage, which can be used to predict ground-level particulate concentration[12,13]. The aerosol optical depth (AOD) is a measure of the extinction of solar radiation by aerosols in the atmospheric column, from the earth's surface to the top of the atmosphere. In contrast, $PM_{2.5}$ is the mass concentration of fine particulate matter measured near the surface. Since both measures (i.e. AOD, $PM_{2.5}$) are affected by the amount of suspended particles in the air, it is commonly assumed that a correlation between AOD and $PM_{2.5}$ can be established and that AOD can be used to predict ground-level $PM_{2.5}$ concentrations after accounting for factors that may interfere with the relationship, e.g. time-varying parameters (RH, temperature, wind speed, etc.). Satellite-retrieved AOD has been widely used to predict ground-level $PM_{2.5}$ concentrations, especially over the areas where ground monitoring stations are not available. For example, in the last decade, satellite-retrieved AOD from different satellite-borne sensors has been used for predicting ground-level $PM_{2.5}$ at varying spatial resolutions, including instruments onboard Low Earth Orbit (LEO) satellites, such as Moderate Resolution Imaging Spectroradiometer (MODIS)[14–19], Multiangle Imaging SpectroRadiometer (MISR)[9,20–22], and Visible Infrared Imaging Radiometer Suite (VIIRS)[23–25], as well as instruments having a Geostationary Earth Orbit (GEO; with only local coverage) satellites such as Himawari[26,27], and GOES[28,29]. In parallel, a new operational MODIS aerosol retrieval algorithm named MultiAngle Implementation of Atmospheric Correction (MAIAC) has been gaining attention as it provides AOD retrievals at very high spatial resolution (1km grid) with global coverage and with each instrument possessing a daily revisit period. MAIAC AOD retrievals have high capability in identifying fine aerosols, emission sources, and aerosol hotspots[30–32]. Hence, it has been widely studied and found to be a powerful predictor of ground-level $PM_{2.5}$ concentrations compared to other AOD products with coarser resolution[33–40]. However, several factors such as meteorology and aerosol types can influence the relationship between AOD and $PM_{2.5}$[41]. Several studies suggested considering other influential factors (such as meteorological variables, land use parameters, and aerosol types) in the AOD-PM modeling to improve the $PM_{2.5}$ prediction using AOD measurements[42 35,39].

Various statistical models have been explored to establish the relation between satellite-retrieved AOD and ground-level $PM_{2.5}$, extending from simple multivariate regression models[13] to more advanced statistical models such as linear mixed effect models[14,15,37,43–45], geographically weighted regression models[46–48], generalized additive models[22,49], and other nonlinear models[49–51]. Moreover, some studies used multiple-stage models to address the spatiotemporal variations in the AOD-$PM_{2.5}$ relationship for more accurate $PM_{2.5}$ predictions[34,39,52].

Recently, machine learning algorithms have been also applied to predict ground-level $PM_{2.5}$ concentrations[27,53–55]. Unlike traditional statistical models, machine learning algorithms have the ability to use a large number of predictors with a few prior assumptions, thus enhancing their predictive power and enabling to capture the complexity in the AOD-$PM_{2.5}$ relationships[56]. Ensemble models such as Random Forest (RF) and the Gradient Boosting (GB) models combine weak learners (multiple models) to obtain more accurate and robust models[57]. RF models have been successfully used to predict $PM_{2.5}$ over several regions, such as in China[53], USA[55] and Italy[40]. In India, very few studies have been conducted to predict ground-level $PM_{2.5}$ concentration using satellite AOD data. For example, Dey et al.[9] and Chowdhury et al.[58] have been used AOD data obtained from MISR and MAIAC AOD respectively, to predict $PM_{2.5}$ concentration by multiplying the AOD with conversion factor obtained from GEOS-Chem chemical transport model. Recently, Mandal et al.[59] implemented multiple-stage modeling including statistical model and machine learning algorithm to predict $PM_{2.5}$ in the national capital of India using satellite data, and land use, meteorological data, and population. However, to the best of our knowledge, no study that reports $PM_{2.5}$ prediction has been conducted across South Asia using an advanced statistical or machine learning model for $PM_{2.5}$ prediction on a regional scale.

In this study, both a statistical model (LME) and a machine learning algorithm (RF) were used to predict, for the first time, high spatially resolved (1 km) ground-level $PM_{2.5}$ concentrations over the IGP region, India; with the MAIAC AOD as an independent variable. The main objective of this study was to examine how accurate can a machine learning model that uses the above satellite-based AOD product be for estimating ground-level $PM_{2.5}$ concentrations. In response to this task, we first, compared the performance of the RF model against that of an LME model, and the more accurate model was used for $PM_{2.5}$ prediction. Next, we have studied the spatiotemporal variation of the estimated $PM_{2.5}$ across the IGP region and identified regional hotspots. The dataset and model details used are described in section 2, the results are presented in section 3, and followed by a discussion in section 4.

2. Data and Method

2.1 Study region

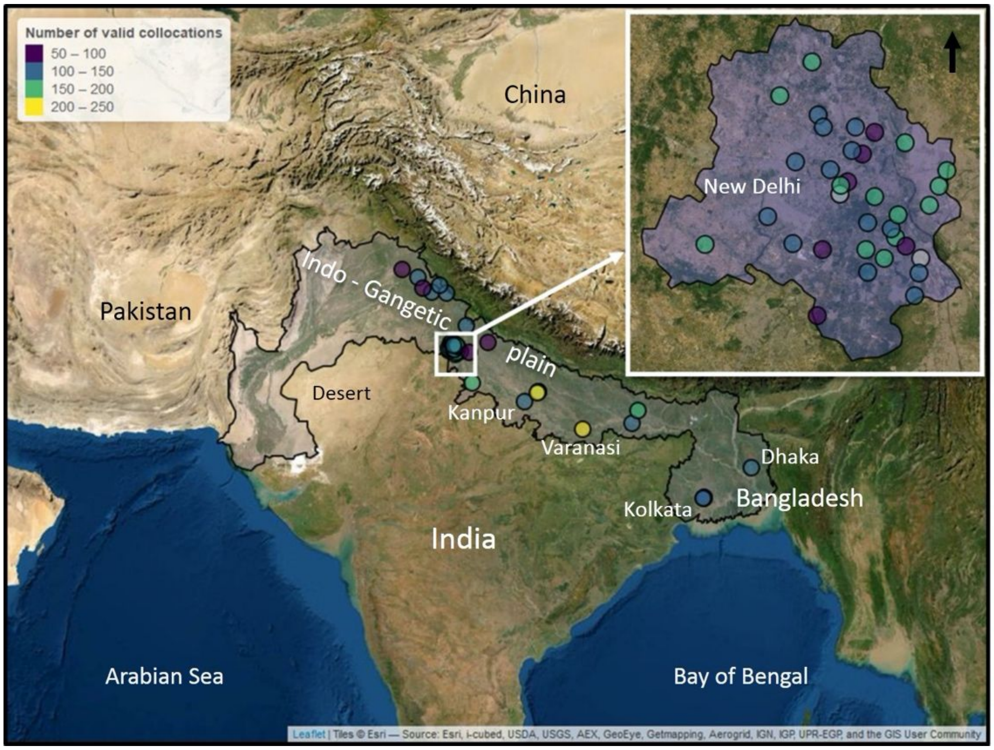


Figure 1: Map of the study region and the spatial distribution of $PM_{2.5}$ monitoring stations. The colors represent the number of $PM_{2.5}$ and independent variables collocations. The shaded area represents the IGP. The area within the box represents the monitoring stations in Delhi.

The study area covers the IGP, which stretches west from Pakistan across Northern India to the east of the Bay of Bengal and Bangladesh (Fig. 1). The IGP region is densely populated and accommodates nearly 13% (>800 million) of the world population. The rapid economic and population growth across the region is associated with a wide range of anthropogenic activities, including biomass/-waste burning, industries, and vehicular emissions, resulting in significant particulate matter pollution across the region. The region is considered to be one of the aerosol hotspots and is characterized by a persistent high aerosol loading throughout the year[60–62].

2.2 Ground-based $PM_{2.5}$ Measurements

Daily mean $PM_{2.5}$ concentrations were obtained from a total of 64 air quality monitoring stations across the IGP from July 1st, 2018, to June 30th, 2019. Specifically, $PM_{2.5}$ data were obtained from 61 monitoring stations of the Central Pollution Control Board (CPCB) (https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing) and 3 $PM_{2.5}$ monitoring stations operated by the US Consulate in Delhi, Kolkata, and Dhaka. The spatial distribution of the air quality monitoring stations across the region is sparse and varying. For example, almost 50% of the monitoring stations are located in New Delhi, with the rest distributed across the major cities. None of the air quality monitoring stations were in rural areas.

Measurement of $PM_{2.5}$ concentrations at all the monitoring stations is done with beta gauge attenuation monitors (BAM-1020; Met One Instruments) that report hourly mean $PM_{2.5}$ concentrations. We calculated the daily mean $PM_{2.5}$ concentrations after applying strict quality control procedures to remove abnormal observation. Only days with more than 14 hourly measurements (60%) were used in the analysis. The geographical location of the monitoring stations and the data availability of the valid MAIAC AOD and $PM_{2.5}$ collocations are shown in Figure 1.

2.3 MODIS MAIAC Products

MAIAC is a relatively new operational MODIS-based aerosol retrieval algorithm that retrieves aerosol properties and columnar water vapor at 1 km spatial resolution over land surface except for snow and ice[32]. The MAIAC aerosol products have higher spatial resolution compared to other operational MODIS aerosol products based on the Dark Target[63] (DT) and the Deep Blue[64] (DB) algorithms. Several validation studies showed that the MAIAC algorithm improves aerosol retrieval accuracy, especially over bright surfaces such as urban areas and dry land[30,32]. Several reasons make MAIAC significantly superior over other operational MODIS algorithms: (a) the high spatial resolution (1km) compared to DT and DB (10km and 3km) that allows to distinguish fine spatial features and to enhance spatial coverage[30], (b) high retrieval accuracy over both dark and bright surfaces[30], and (c) MAIAC'S capability to retrieve AOD for different aerosol types while discriminating among absorbing fine (smoke) and coarse (dust) aerosols[32].

In this study, the combined Terra and Aqua MAIAC product (MCD19A2; https://ladsweb.modaps.eosdis.nasa.gov/) was used to extract Terra and Aqua AOD at 550nm with an aerosol type (compositional) label (dust, smoke, and background), and CWV. Only the highest quality data, designated with the Quality Assurance (QA) cloud mask value "clear", were used.

The spatial coverage of Terra (~10:30 am overpass time) and Aqua (~01:30 pm overpass time) MAIAC AOD varies due to the diurnal cycle of cloud cover[30], meteorological conditions (mainly the lower atmospheric boundary layer[65]) and the daily varying anthropogenic activities[66]. Therefore, a combined MAIAC AOD product from both Terra and Aqua, can enhance the spatial and temporal coverage and provide a more representative AOD that accounts for both the morning (Terra) and afternoon (Aqua) time windows, from ~10:00 am until 02:00 pm local time. Nevertheless, if one of the two values (Aqua or Terra AOD) is missing the combined AOD product will be biased towards either the morning or the afternoon retrieval. To eliminate this bias, missing Terra AOD were predicted from Aqua AOD, and vice versa, by fitting seasonal linear regression models to both the Aqua and Terra AOD. Table S1 shows the seasonal regression equations and correlation coefficients of each season for both the AOD and CWV. The number of the available combined AOD product increased by 22% and 24% compared to Terra or Aqua only AOD retrievals, respectively. The R^2 of the seasonal regressions between Aqua and Terra MAIAC AOD ranged from 0.63-0.79 ($p < 0.001$).

2.4 Meteorological data

Meteorological variables, including the ambient temperature at 2 m a.g.l. (temp; K), surface pressure (SF; hPa), wind field at 10 m a.g.l. (Wind Speed (WS); m s^{-1}, and Wind Direction (WD); ^{o}), relative humidity (RH; %), and the planetary boundary layer height (PBLH; m), were obtained from the European Center for Medium-Range Forecast (ECMWF) atmospheric reanalysis ERA-Interim products (https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/). The spatial resolution of ERA-Interim is 12.5km, and its temporal resolution is 6h except for the PBLH that is provided in every 3h. All the meteorological variables were averaged over the time window corresponding to the Terra and Aqua overpass times.

2.5 Auxiliary data

Level 3 Terra and Aqua MODIS 16-day composite Normalized Difference Vegetation Index (NDVI) data (MxD13A2, x is O or Y for Terra and Aqua, respectively) at 1km spatial resolution were used in this study as a proxy for the land use parameter. The NDVI data are reported every 16 days for both Terra and Aqua but with 8 days of difference between them (Terra reports on day 001 while Aqua reports on day 009). This corresponds to having a measurement in every 8 days. Elevation (Elev) data were obtained from the Shuttle Radar Topography Mission (SRTM) database (http://srtm.csi.cgiar.org/srtmdata/) at 30m spatial resolution[67] and used as spatial predictors.

2.6 Data processing and integration

For predicting $PM_{2.5}$ concentration at 1km spatial resolution using MAIAC AOD and other temporal and spatial predictors, the spatial resolution of all the predictors should be consistent and matched with the MAIAC AOD grid. Therefore, all the meteorological data and auxiliary data were re-projected and gridded to match the MAIAC AOD fixed grid. In particular, the Terra and Aqua MODIS NDVI were gridded to a 1km, and then the combined Terra and Aqua daily NDVI was calculated using the temporally interpolated spline function technique. The meteorological data were gridded to a 1km grid using bi-linear interpolation and temporal subsets that matched the Terra and Aqua overpass times (~10 am–2 pm local time). Similarly, elevation data were also gridded to 1km. The total number of spatial and temporal $PM_{2.5}$, satellite- and meteorological data was 8233 matched up collocations, distributed over 356 days from July 1st, 2018 to June 30th, 2019.

2.7 Model Development

2.7.1 Linear Mixed Effect (LME) Model

Linear mixed effect (LME) models have been widely used to estimate $PM_{2.5}$ concentrations based on satellite-derived AOD[68] since it controls for the inherent day-to-day variability in the relationship between AOD and $PM_{2.5}$. The AOD-$PM_{2.5}$ relationship is expected to be influenced by time-varying parameters such as RH, PBLH, Temp, and the optical properties of the particles and their vertical distribution[41]. Therefore, considering the daily variability in the AOD-$PM_{2.5}$ relationship is essential to improve the correlation between the AOD and $PM_{2.5}$. Hence, allowing for a day-specific random slope and intercept enables to examine the day-to-day variability in the AOD- $PM_{2.5}$ relationship.

In this study, we developed a nested day- and month-specific random effect based on all the days with valid AOD-$PM_{2.5}$ collocations (the days with less than three collocations were removed from the dataset). The LME model structure is expressed by the following equation (Eq. 1):

$$PM_{ij} = (\alpha_0 + (\alpha_{day} + \alpha_{month})) + (\beta_0 + (\beta_{day} + \beta_{month})) \times AOD_{ij} + \beta_1\, CWV_{ij} + \beta_2\, WS_{ij} + \beta_3\, RH_{ij} + \beta_4\, PBLH_{ij} + \beta_5\, SP_{ij} + \beta_6\, WD_{ij} + \beta_7\, NDVI_{ij} + \beta_8\, Temp_{ij} + \beta_9\, Elev_{ij} + \varepsilon_{ij} \qquad \text{Eq. (1)}$$

where PM_{ij} and AOD_{ij} are the $PM_{2.5}$ concentration and MAIAC AOD at monitoring site *i* on day *j*; α_0 and β_0 are the fixed intercept and slope, respectively; α_{day} , α_{Month}, β_{day} and β_{month} are the day- and month-specific random intercept and slope, respectively; CWV, WS, WD, RH, PBLH, SP, NDVI, Temp, and Elev are the corresponding auxiliary variables at grid *i* and day *j* (and their corresponding fixed slopes); and ε_{ij} is the error term at site *i* and on day *j*.

The day- and month-specific intercepts and slopes allow the model to control day-to-day and monthly variability in the relationship between $PM_{2.5}$ and AOD. Spatial predictors such as NDVI, and Elevation were found to be significantly correlated with the $PM_{2.5}$ and, therefore, were included in the model. All the variables were tested and only the significant ones were used during the model fitting.

2.7.2 Random Forest (RF) model

Random forest is an ensemble learning (algorithm) that aggregates a large number of decision trees which were created independently using the bootstrap resampling method[69]. The bagging (bootstrap aggregation) technique allows to reduce the variance of the estimated prediction by averaging the regression results from all decision trees. Each node of the tree splits into two daughter nodes using the best split from the randomly selected variables[69]. Aggregating weak learners into a strong learner leads to a final model of enhanced performance. Moreover, the random forest model provides an estimate for the importance of each variable by measuring the increase in the prediction error (decrease in the accuracy score) of the final model after performing variable permutations. Here, the mean decrease accuracy is calculated by the permutation scheme of Breiman[70]. In the random forest algorithm, the main two variables that have the major effect at each level (bifurcation) on the model accuracy are mostly the ones used to split the residual subset at each node (mtry) and to select the number of trees in the forest (ntree). In this experiment, we found that the best model accuracy was obtained for mtry=12 and ntree=1500. The unscaled variables importance[71] of the final model is reported in Table S2.

2.8 Evaluation of Models

To evaluate the performance of the developed models across the IGP, we adopted two 10-fold cross-validations (CV) approaches a site-based CV, and a sample-based CV. The 10-fold CV method[72] randomly split the database into ten subsets, each containing 10% of the data. In each round, the model trains on nine subsets (90% of the data) and predicts the 10th subset, with the predictions evaluated against the true data. The process is repeated ten times thus ensuring that every subset has been evaluated. In the site-based CV, the database is split according to the monitoring sites into ten subsets, each containing ~10% of the data. As such, each subset contains different monitoring stations. In each round, one subset is held-out and PM levels at the sites it contains are predicted using the model that has been developed based on data from the other sites. The model is evaluated by comparing its predictions in the held-out sites against the true observations (which have not been used for the model development). This process is repeated with each subset held-out in turn. The site-based CV is used for assessing how well the model performs over regions that do not have monitoring sites, such that the prediction must be done by applying a model that has been developed (and evaluated) over another region. In the sample-based CV, the same procedure is performed using the whole database without accounting for the monitoring site it comes from. As such, the sample-based CV is used for a general assessment of the model performance for filling data gaps in both space and time in regions were monitoring sites do exist. The performance of the CV predictions has been examined using several statistical metrics, including the Root Mean Squared

Error (RMSE), Relative Prediction Error (RPE; Eq. 2), coefficient of determination (R^2), Mean Prediction Error (MPE; Eq. 3), and the slope (*b*) and intercept (*a*) of the linear regression between the predicted and observed $PM_{2.5}$. The RPE and the MPE are calculated as:

$$\mathrm{RPE} = \frac{\mathrm{RMSE}}{\overline{\mathrm{PM}_{2.5}}} \times 100 \qquad \text{Eq. (2)}$$

$$\mathrm{MPE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left|\mathrm{Predicted_i} - \mathrm{Observed}_{\mathrm{i}}\right| \qquad \text{Eq. (3)}$$

3. Results and Discussion

3.1 Descriptive statistics

The histograms and descriptive statistics for all the dependent and independent variables used for the model development are illustrated in Figure S1 and Tables S3. The annual mean $PM_{2.5}$ over the entire region was 114.49 ± 76.65 μg/m³ (*N*=8,233), and the seasonal mean $PM_{2.5}$ were winter: 170.16 ± 88.46 μg/m³, postmonsoon: 150.69 ± 73.16 μg/m³, premonsoon: 77.59 ± 34.38 μg/m³, and monsoon: 58.00 ± 21.98 μg/m³. The overall mean AOD was 0.57 ± 0.39, and the seasonal means were winter: 3.28 ± 0.40, postmonsoon: 0.77 ± 0.55, premonsoon: 0.40 ± 0.20 and monsoon: 3.13 ± 0.28 (Table S3). Notably, while the highest $PM_{2.5}$ was observed in the winter and the lowest $PM_{2.5}$ was observed in the monsoon, the AOD showed much smaller variation with the highest retrievals during the postmonsoon and the lowest retrievals during the pre-monsoon seasons. The seasonal discrepancies between AOD and $PM_{2.5}$, in particular, the low $PM_{2.5}$ concentrations but high AOD values during the monsoon season, are attributed to the abundance in water vapor in the atmospheric column during monsoon (CWV =3.32 ± 0.65), which favors hygroscopic growth of the aerosol particles[30,61]. Hygroscopic growth of aerosol particles enhances scattering, thus resulting in higher AOD[73]. In contrast, $PM_{2.5}$ is measured near the surface at a fixed RH of <40% and does not reflect hygroscopic growth as in the free air. Both the AOD and $PM_{2.5}$ data showed a similar unimodal distribution, with the correlation coefficient between the daily mean $PM_{2.5}$ and the combined Aqua and Terra MAIAC AOD being r =0.47 (p < 0.0001). The variables used in this study, i.e. meteorological variables, boundary layer height, and land use and the land cover attributes, were all significantly correlated (p < 0.0001) with the $PM_{2.5}$ (Table S4).

Furthermore, the variance inflation factors (VIF) was used to quantify the collinearity among the predictors, which could affect the model performance. A VIF value of 10 was set as the threshold for collinearity. All the VIF values were <10, i.e. showing little to nil collinearity (Table S5).

3.2 Models fitting and evaluation

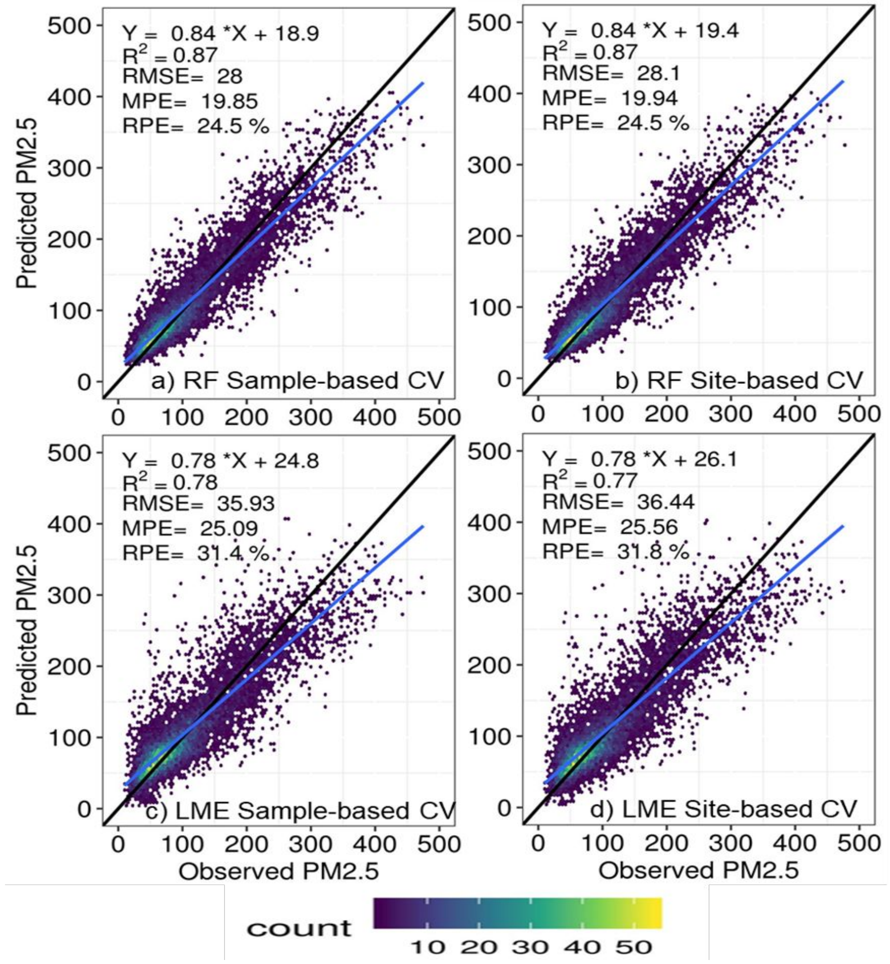


Figure 2: Scatterplot of the cross-validation results. Left column: sample-based cross-validation, right column: site-based cross-validation, upper row: RF model, and lower row: LME model.

Figure 2 shows scatter plots of the sample-based and the site-based CV predicted vs observed daily mean $PM_{2.5}$ for the LME and the RF models. Clearly, the RF model performed better, with R^2 of 0.87 and RMSE of 28 μg/m^3 (irrespective of the CV method applied), compared to the LME model (R^2 ~78%, RMSE ~36 μg/m^3). Both the RF and LME models tend to underestimate the ground-level $PM_{2.5}$ concentrations, especially on highly polluted days ($PM_{2.5}$ >100 μg/m^3), with the underestimation more severe when using the LME model compared to the RF model. Since the sample-based and site-based CV methods resulted in almost identical results, both the LME and RF models were apparently not over-fitted to the data, suggesting a good spatial predictive power. Still, the RF model outperformed the LME in terms of accuracy, having lower RPE (RF: 24.5%, LME: ~ 31.6%).

To evaluate the performance of the RF and LME models at different temporal averaging scales, the weekly, monthly, seasonal, and annual $PM_{2.5}$ means were calculated based on daily predictions from days for which >20% of the site-specific daily $PM_{2.5}$ predictions were available (figure S2 and Table S6). Like on the daily scale, the RF model was more accurate than the LME model also on the weekly, monthly, seasonal, and annual scales, with high R^2 (0.91-0.92), a slope close to unity (0.88-0.9), and a lower RPE (monthly: 15.1%, seasonal: 13.9%, annual: 8.8%).

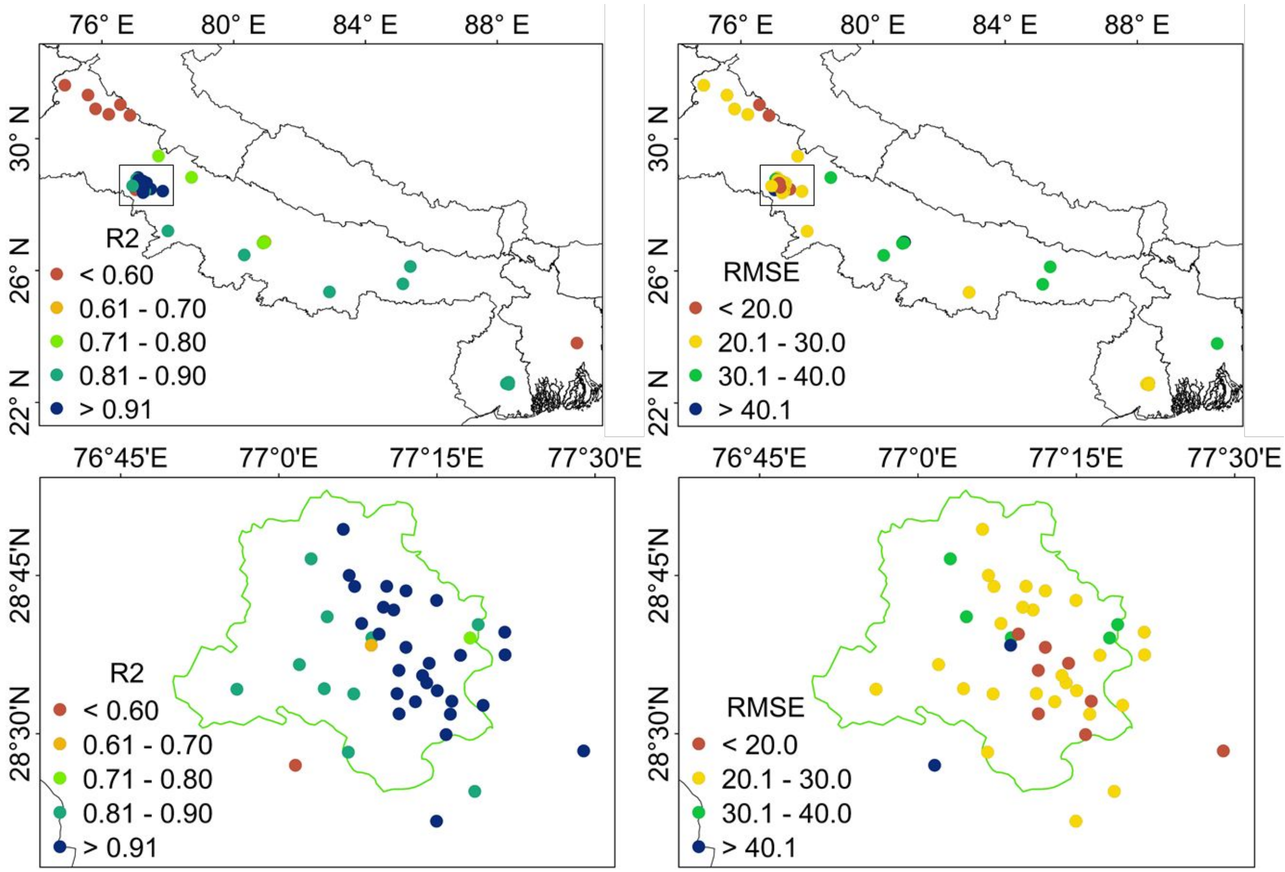


Figure 3: Spatial distribution of the R^2 (left panels) and RMSE (right panels) between the observed and predicted $PM_{2.5}$ using the RF model. Upper row: across the whole IGP region, lower row: the Delhi area.

Figure 3 shows the spatial distribution of the site-based CV performance metrics of the RF model, with Fig. 3(a, b) depicting the R^2 and RMSE across the IGP, respectively, and Fig. 3(c, d) focusing on the Delhi city. The overall IGP mean R^2 was 0.81, with 85% of the stations showing R^2 >0.7. The lower R^2 (<0.6) were found at the northwest IGP, which may be attributed to the small number of data points (collocations) in this region due to limited $PM_{2.5}$ (only a few months), and may not represent the entire year. Similarly, while the IGP average RMSE was 26.7 µg/m^3 a relatively high RMSE values (>30 µg/m^3) were evident in few stations in the middle and lower IGP, attributed to the high $PM_{2.5}$ levels throughout the year (annual average >150 µg/m^3). Since the model was trained with around 50% of the data obtained from stations in Delhi, therefore the performance metrics were relatively better than the stations located in upper and lower IGP. Overall, the RF model achieved satisfactory performance and was able to capture most of the variability in the $PM_{2.5}$ across the region, with R^2 >0.7 in most of the monitoring stations.

The seasonal variation in aerosol sources and meteorological variables also affected the AOD-PM model performance seasonally. In premonsoon and monsoon seasons, the IGP is affected by aerosols transported by the southwest monsoon and frequently associated with higher wind speed and deeper boundary layer. While in winter, the $PM_{2.5}$ primarily concentrates near the surface due to shallow boundary layer and slower wind speed[61,62]. Model performances both in cold seasons i.e., winter (RPE: 20.9%) and postmonsoon (RPE: 22.3%) was also compared with warm seasons including premonsoon (RPE: 28%) and monsoon (RPE: 26.5%) and shown in Table S7. The larger slope of the fitting line for colder seasons reflect a higher $PM_{2.5}$ that was concentrated near the surface due to a shallow PBL (788-1113m) compared to the warmer seasons when the PBLH was relatively higher (1673-1901m)[74].

3.3 Predicted $PM_{2.5}$ over IGP

Figure 4 shows the annual mean satellite-based $PM_{2.5}$ estimates for IGP at 1 km grid resolution, as derived from the RF model. The overall estimated annual mean $PM_{2.5}$ (July 1st, 2018 to June 30th, 2019) was 112.7 µg/m^3, which exceeds the 40 µg/m^3 Indian National Ambient Air Quality Standards (NAAQS). In particular, the middle and lower IGP regions experience higher $PM_{2.5}$ concentrations (>110 µg/m^3), with around 79.3% of the area experiencing an annual mean $PM_{2.5}$ concentration between 110-150 µg/m^3. The highest annual mean $PM_{2.5}$ was found over the state of Bihar, West of Bengal, and Bangladesh, with $PM_{2.5}$ concentrations exceeding 130 µg/m^3. The high $PM_{2.5}$ levels in the middle and lower IGP are most likely due to the combined contributions of local sources and long-range transport from the upper IGP[61].

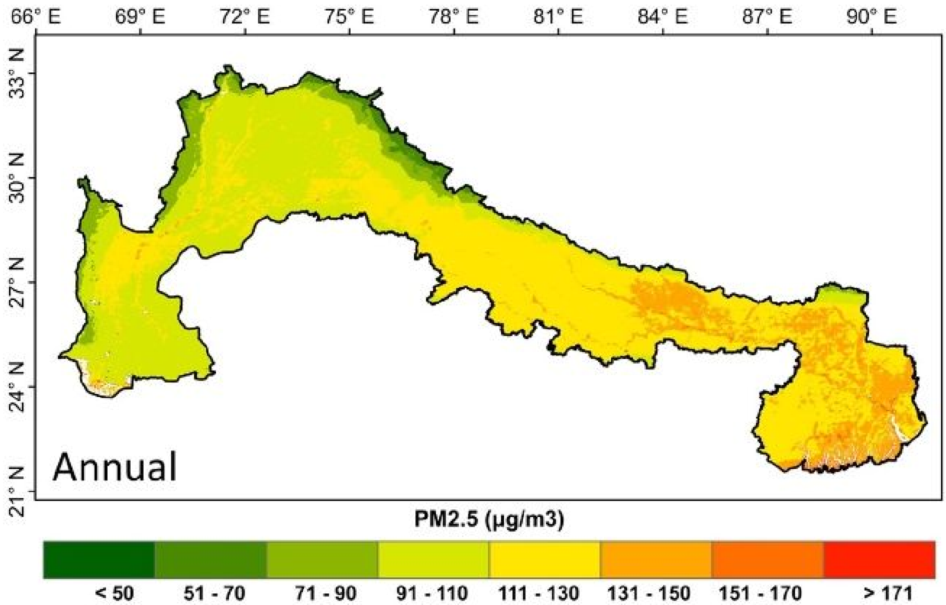


Figure 4: Spatial distribution of annual mean $PM_{2.5}$ estimates at 1 km grid resolution over IGP.

Seasonally, a significant variation is noted across the region with the highest $PM_{2.5}$ levels recorded in winter (DJF) (154 ± 22.4 µg/m^3) (Fig. S3). Spatial differences are also evident, with about 66% of the IGP exposed to $PM_{2.5}$ concentrations >150 µg/m^3 in the winter, and 25% of the IGP region (mainly the middle and lower IGP) experiencing $PM_{2.5}$ >170 µg/m^3. High $PM_{2.5}$ levels are also estimated during postmonsoon (ON; 128.8 ± 16.0 µg/m^3), with about 87% of the IGP exposed to $PM_{2.5}$ in the range 110 - 150 µg/m^3. The lowest $PM_{2.5}$ levels were estimated in the monsoon and premonsoon seasons, with mean $PM_{2.5}$, mean of 59.9 ± 5.9 µg/m^3 and 80.9 ± 9.5 µg/m^3, respectively.

Taking Delhi as an example for one of the most heavily $PM_{2.5}$-polluted metropolitans/megapolises in South Asia and the world, we also examined the capability of our model to capture $PM_{2.5}$ variability at the urban scale. The true color image (Fig. 5a) and the annual mean $PM_{2.5}$ estimates over Delhi (Figure 5b) show high $PM_{2.5}$ concentration in central and eastern Delhi – the most densely populated areas, and lower levels in southern Delhi; which is greener and not as densely populated, with an overall annual mean of $PM_{2.5}$ of 121.8 ± 7.4 µg/m^3, i.e. 8% higher than the IGP mean $PM_{2.5}$. These results suggest that both local particulate sources combined with long-range transport of aerosol from the north-west IGP, especially during stubble burning period [61, 75,76], could be captured by the model, which accounts for enhanced $PM_{2.5}$ concentrations in Delhi.

To critically examine a severe $PM_{2.5}$ condition, we selected the stubble burning episode, which occurs every year in November in the Punjab and Haryana states, and affects the whole northern India[76]. Figure 5(c, d) shows the spatial distribution of active fires on November 8th, 2018, obtained from the VIIRS and MODIS (Aqua and Terra) sensors (https://firms.modaps.eosdis.nasa.gov/) together with the estimated $PM_{2.5}$. Clearly, areas located downwind of the fire spots experienced higher $PM_{2.5}$ and the model shows good sensitivity for capturing these high $PM_{2.5}$ areas, demonstrating the excellent capability of the model to identify pollution sources in both space and time. Examining the model performance at different Indian air quality categories, for $PM_{2.5}$ >60 μg/m^3 (moderate air quality) the CV R^2 was 0.84 and the RPE was 21.8%, for $PM_{2.5}$ >90 μg/m^3 the R^2 was 0.79 and the RPE was 19.8%, and for $PM_{2.5}$ >120 μg/m^3 (very poor air quality) the R^2 and RPE were 0.71 and 20.82%, respectively (Fig. S4). The model performance at low $PM_{2.5}$ (cleaner conditions) was poorer than at the polluted conditions (Fig. S4).

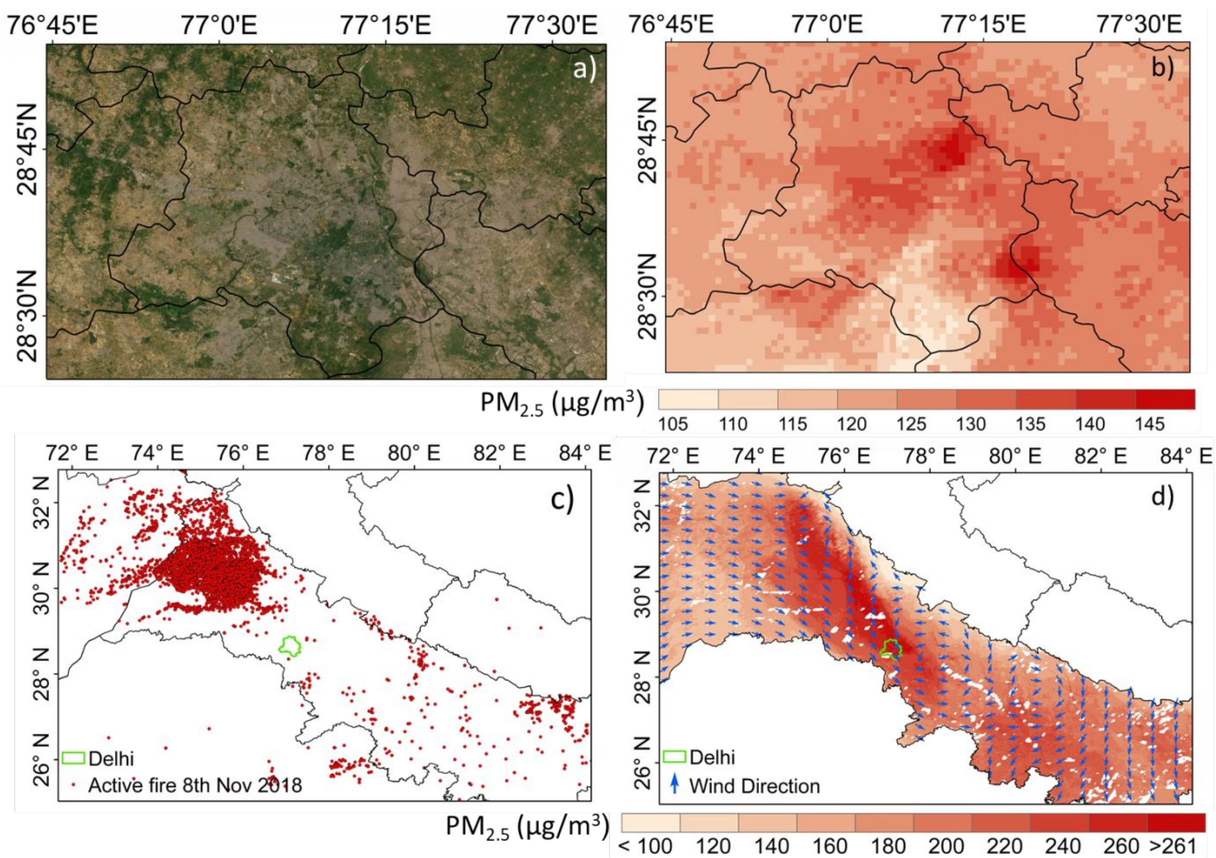


Figure 5: (a) RGB image, b) annual mean $PM_{2.5}$ over Delhi, c) Active fire counts in the northwest IGP obtained from VIIRS and MODIS (Aqua and Terra) sensors on November 8th, 2018, and d) $PM_{2.5}$ estimates superimposed by the wind direction during the same day.

4. Discussion

RF and LME models were developed to predict daily ground-level $PM_{2.5}$ concentrations across the IGP, South Asia. A few studies have estimated $PM_{2.5}$ in the IGP region based on satellite retrieved AOD. Dey et al.[9] used AOD retrievals from the MISR sensor to estimate ground-level $PM_{2.5}$ at a spatial resolution of 0.5°

x 0.5˚, using a scale factor obtained from the GEOS-Chem chemical transport model. Similarly, using a scale factor from GEOM-Chem, Chowdhury et al.[58] estimated ground-level $PM_{2.5}$ concentrations during the dry season (October–June) at a spatial resolution of 1km grid over the Delhi National Capital Region (NCR). Similarly, over Delhi, Mandal et al.[59] estimated the $PM_{2.5}$ from 2010 to 2016 at 1km spatial grid using the multi-stage prediction model. To the best of our knowledge, the current study is the first regional-scale study in South Asia to predict daily $PM_{2.5}$ at a high spatial resolution (1km x 1km), using satellite retrievals of AOD and CWV, together with meteorological and land use information, and applying both the random forest (machine learning) algorithm as well as an advanced statistical model (LME). To date, only few attempts were made to use satellite-based AOD for estimating ground-level $PM_{2.5}$ across South Asia, as the region is severely constrained by the availability of quality surface monitoring data which are essential for model calibration and validation. Taking advantage of the recently established air quality monitoring network, we developed regional-scale models for estimating daily $PM_{2.5}$ concentrations over the IGP. The RF model exhibited adequate performance and higher accuracy than the LME model, with better cross-validated explained variance (R^2 =0.87) and lower prediction error (RPE =24.5%). Our RF model performed similarly to- or better than previous RF models that were developed for China (R^2= 0.83-0.85, RPE=30.7%-35.9%)[53,54] and the USA (R^2=0.80, RPE=29.2%)[55]. The model showed satisfactory predictive capability across the region with comparable site-based CV and sample-based CV results. Moreover, the RF model also revealed high accuracy in estimating weekly (R^2 =0.91, RPE =17.7%), monthly (R^2= 0.92, RPE= 15.5%), seasonal (R^2 =0.92, RPE =13.91%), and annual (R^2 =0.90, RPE =8.8%) mean $PM_{2.5}$ levels. The high spatial resolution and low-bias of the $PM_{2.5}$ estimates (both weekly and monthly mean) support using it in different research domains, especially in environmental epidemiology and climatological studies.

Due to the lack of historical $PM_{2.5}$ records across the IGP, the year-to-year variability in $PM_{2.5}$ concentrations cannot be assessed. Similarly, the insufficient number of observations and the low $PM_{2.5}$ concentrations in the northwestern IGP resulted in poor model performance compared to other parts of IGP. Indeed, the model results in greater accuracy when high $PM_{2.5}$ concentrations were experienced.

The modeled $PM_{2.5}$ map showed significant spatial and temporal variation across the IGP. Seasonally, winter and postmonsoon are the more polluted seasons while the wet monsoon season is the cleaner one. Anthropogenic activities such as open burning stubble during postmonsoon, and burning of biomass and coal for heating and cooking, combined with shallow atmospheric boundary layer height, lead to enhanced $PM_{2.5}$ concentrations during the winter and post-monsoon. The lower $PM_{2.5}$ concentrations in the monsoon period are due to wet deposition, strong convection, and higher boundary layer heights. Nonetheless, the $PM_{2.5}$ concentrations in all seasons were higher than the Indian NAAQS (annual average: 40 μg/m^3).

Spatially, the middle and lower IGP showed poor air quality compared to the upper IGP. In winter, the middle and lower IGP experience very poor air quality, with mean $PM_{2.5}$ concentrations >170 μg/m^3. The highly spatially resolved $PM_{2.5}$ estimates were found to have potential to identify $PM_{2.5}$ hotspots and to study $PM_{2.5}$ on small scales, especially in urban areas. Our model performed well at the urban scale, showing the good capability to capture spatial $PM_{2.5}$ variability.

Finally, the random forest machine learning algorithm showed high skill in predicting $PM_{2.5}$ by fusing satellite aerosol products, meteorological models' output, and land use data. Future improvements of the model may involve using richer land use parameters (i.e. the road network, vehicle volumes) and emissions data (agricultural residues burning, industries emissions inventory, municipal solid waste burning, etc.) which may be helpful to further improve the reliability of the AOD-PM model across the Indo-Gangetic plain.

5. Acknowledgment

The research is supported by the India-Israel bilateral research grant funded by the University Grants Commission to TB (Grant No. 6-11/ 2018 IC) and the Israel Science Foundation to DB (Grant No. 0472714). TB also acknowledges funds received from ASEAN- India S&T Development Fund, Govt. of India (CRD/2018/000011) under ASEAN- India Collaborative Research and Development Scheme. AM acknowledges USRA, NAMS R&D Student Program, and Jawaharlal Nehru Scholarship for Doctoral studies from Jawaharlal Nehru Memorial Fund; MSH acknowledges the NASA Post-Doctoral Fellowship administered by USRA.

Data availability

MODIS MAIAC data is available at https://ladsweb.modaps.eosdis.nasa.gov. Modis Fire products are obtained from Fire Information for Resource Management System (FIRMS) (https://firms.modaps.eosdis.nasa.gov). All datasets were last accessed in November 2019.

Author Contributions

AM: methodology, formal analysis, review and writing draft manuscript; **MSH, MB, AL, RC, DB**: methodology and interpretation; review and editing draft; **TB**: methodology and interpretation, resources, review, writing and editing draft manuscript.

Competing interests. Authors declare that they have no conflict of interest.

Supporting Information. The supporting figures and tables are included in supplementary file.

6. Bibliography

(1) GBD 2013 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **2015**, *386*, 2287–2323.

(2) Hayes, R.B., Lim, C., Zhang, Y., Cromar, K., Shao, Y., Reynolds, H.R., Silverman, D.T., Jones, R.R., Park, Y., Jerrett, M. and Ahn, J., PM2.5 air pollution and cause-specific cardiovascular disease mortality. Int. J. Epidemiol. 2020, 49, 25–35.

(3) Chowdhury, S.; Dey, S.; Smith, K. R. Ambient PM2.5 exposure and expected premature mortality to 2100 in India under climate change scenarios. *Nat. Commun.* **2018**, *9*, 318.

(4) Brook, R. D.; Newby, D. E.; Rajagopalan, S. Air pollution and cardiometabolic disease: an update and call for clinical trials. *Am. J. Hypertens.* **2017**, *31*, 1–10.

(5) Weber, S. A.; Insaf, T. Z.; Hall, E. S.; Talbot, T. O.; Huff, A. K. Assessing the impact of fine particulate matter (PM2.5) on respiratory-cardiovascular chronic diseases in the New York City Metropolitan area using Hierarchical Bayesian Model estimates. *Environ. Res.* **2016**, *151*, 399–409.

(6) Thurston, G. D.; Ahn, J.; Cromar, K. R.; Shao, Y.; Reynolds, H. R.; Jerrett, M.; Lim, C. C.; Shanley, R.; Park, Y.; Hayes, R. B. Ambient Particulate Matter Air Pollution Exposure and Mortality in the NIH-AARP Diet and Health Cohort. *Environ. Health Perspect.* **2016**, *124*, 484–490.

(7) Lelieveld, J.; Evans, J. S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **2015**, *525*, 367–371.

(8) World Health Organization. *Global action plan on physical activity 2018–2030: more active people for a healthier world. Geneva:World Health Organization; 2018. Licence: CC BY-NC-SA 3.0 IGO.* ; 2018.

(9) Dey, S.; Di Girolamo, L.; van Donkelaar, A.; Tripathi, S. N.; Gupta, T.; Mohan, M. Variability of outdoor fine particulate (PM2.5) concentration in the Indian Subcontinent: A remote sensing approach. *Remote Sens. Environ.* **2012**, *127*, 153–161.

(10) Bilal, M., Nichol, J.E., Nazeer, M., Shi, Y., Wang, L., Kumar, K.R., Ho, H.C., Mazhar, U., Bleiweiss, M.P., Qiu, Z. and Khedher, K.M. Characteristics of Fine Particulate Matter (PM2.5) over Urban, Suburban, and Rural Areas of Hong Kong. Atmosphere 2019, 10, 496..

(11) Murari, V.; Singh, N.; Ranjan, R.; Singh, R. S.; Banerjee, T. Source apportionment and health risk assessment of airborne particulates over central Indo-Gangetic Plain. *Chemosphere* **2020**, 127145.

(12) Wang, J. Intercomparison between satellite-derived aerosol optical thickness and $PM_{2.5}$ mass: Implications for air quality studies. *Geophys. Res. Lett.* **2003**, *30*, 2095.

(13) Gupta, P.; Christopher, S. A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. *J. Geophys. Res.* **2009**, *114*.

(14) Xiao, Q.; Wang, Y.; Chang, H. H.; Meng, X.; Geng, G.; Lyapustin, A.; Liu, Y. Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sens. Environ.* **2017**, *199*, 437–446.

(15) Sorek-Hamer, M.; Kloog, I.; Koutrakis, P.; Strawa, A. W.; Chatfield, R.; Cohen, A.; Ridgway, W. L.; Broday, D. M. Assessment of PM 2.5 concentrations over bright surfaces using MODIS satellite observations. *Remote Sens. Environ.* **2015**, *163*, 180–185.

(16) Chudnovsky, A. A.; Koutrakis, P.; Kloog, I.; Melly, S.; Nordio, F.; Lyapustin, A.; Wang, Y.; Schwartz, J. Fine particulate matter predictions using high resolution Aerosol Optical Depth (AOD) retrievals. *Atmos. Environ.* **2014**, *89*, 189–198.

(17) Zou, B.; Pu, Q.; Bilal, M.; Weng, Q.; Zhai, L.; Nichol, J. E. High-Resolution Satellite Mapping of Fine Particulates Based on Geographically Weighted Regression. *IEEE Geosci. Remote Sensing Lett.* **2016**, *13*, 495–499.

(18) Bilal, M.; Nichol, J. E.; Spak, S. N. A new approach for estimation of fine particulate concentrations using satellite aerosol optical depth and binning of meteorological variables. *Aerosol Air Qual. Res.* **2017**, *17*, 356–367.

(19) Chu, H.-J.; Bilal, M. PM2.5 mapping using integrated geographically temporally weighted regression (GTWR) and random sample consensus (RANSAC) models. *Environ. Sci. Pollut. Res. Int.* **2019**, *26*, 1902–1910.

(20) Franklin, M.; Kalashnikova, O. V.; Garay, M. J. Size-resolved particulate matter concentrations derived from 4.4 km-resolution size-fractionated Multi-angle Imaging SpectroRadiometer (MISR) aerosol optical depth over Southern California. *Remote Sens. Environ.* **2017**, *196*, 312–323.

(21) Franklin, M.; Chau, K.; Kalashnikova, O.; Garay, M.; Enebish, T.; Sorek-Hamer, M. Using Multi-Angle Imaging SpectroRadiometer Aerosol Mixture Properties for Air Quality Assessment in Mongolia. *Remote Sens (Basel)* **2018**, *10*, 1317.

(22) Meng, X.; Garay, M. J.; Diner, D. J.; Kalashnikova, O. V.; Xu, J.; Liu, Y. Estimating PM 2.5' ' speciation concentrations using prototype 4.4 km-resolution MISR aerosol properties over Southern California. *Atmos. Environ.* **2018**, *181*, 70–81.

(23) Wu, J.; Yao, F.; Li, W.; Si, M. VIIRS-based remote sensing estimation of ground-level PM 2.5 concentrations in Beijing–Tianjin–Hebei: A spatiotemporal statistical model. *Remote Sens. Environ.* **2016**, *184*, 316–328.

(24) Yao, F.; Si, M.; Li, W.; Wu, J. A multidimensional comparison between MODIS and VIIRS AOD in estimating ground-level PM2.5 concentrations over a heavily polluted region in China. *Sci. Total Environ.* **2018**, *618*, 819–828.

(25) Yao, F.; Wu, J.; Li, W.; Peng, J. Estimating Daily PM2.5 Concentrations in Beijing Using 750-M VIIRS IP AOD Retrievals and a Nested Spatiotemporal Statistical Model. *Remote Sens (Basel)* **2019**, *11*, 841.

(26) Wang, W.; Mao, F.; Du, L.; Pan, Z.; Gong, W.; Fang, S. Deriving Hourly PM2.5 Concentrations from Himawari-8 AODs over Beijing–Tianjin–Hebei in China. *Remote Sens (Basel)* **2017**, *9*, 858.

(27) Zang, L.; Mao, F.; Guo, J.; Wang, W.; Pan, Z.; Shen, H.; Zhu, B.; Wang, Z. Estimation of spatiotemporal PM1.0 distributions in China by combining PM2.5 observations with satellite aerosol optical depth. *Sci. Total Environ.* **2019**, *658*, 1256–1264.

(28) Reid, C. E.; Jerrett, M.; Petersen, M. L.; Pfister, G. G.; Morefield, P. E.; Tager, I. B.; Raffuse, S. M.; Balmes, J. R. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environ. Sci. Technol.* **2015**, *49*, 3887–3896.

(29) Liu, Y.; Paciorek, C. J.; Koutrakis, P. Estimating regional spatial and temporal variability of PM(2.5) concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.* **2009**, *117*, 886–892.

(30) Mhawish, A.; Banerjee, T.; Sorek-Hamer, M.; Lyapustin, A.; Broday, D. M.; Chatfield, R. Comparison and evaluation of MODIS Multi-angle Implementation of Atmospheric Correction (MAIAC) aerosol product over South Asia. *Remote Sens. Environ.* **2019**, *224*, 12–28.

(31) Jethva, H.; Torres, O.; Yoshida, Y. Accuracy assessment of MODIS land aerosol optical thickness algorithms using AERONET measurements over North America. *Atmos. Meas. Tech.* **2019**, *12*, 4291–4307.

(32) Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741–5765.

(33) Hu, X.; Waller, L. A.; Lyapustin, A.; Wang, Y.; Al-Hamdan, M. Z.; Crosson, W. L.; Estes, M. G.; Estes, S. M.; Quattrochi, D. A.; Puttaswamy, S. J. and Liu, Y. Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sens. Environ.* **2014**, *140*, 220–232.

(34) Kloog, I.; Chudnovsky, A. A.; Just, A. C.; Nordio, F.; Koutrakis, P.; Coull, B. A.; Lyapustin, A.; Wang, Y.; Schwartz, J. A New Hybrid Spatio-Temporal Model For Estimating Daily Multi-Year PM2.5 Concentrations Across Northeastern USA Using High Resolution Aerosol Optical Depth Data. *Atmos. Environ.* **2014**, *95*, 581–590.

(35) Kloog, I.; Sorek-Hamer, M.; Lyapustin, A.; Coull, B.; Wang, Y.; Just, A. C.; Schwartz, J.; Broday, D. M. Estimating daily PM 2.5 and PM 10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. *Atmos. Environ.* **2015**, *122*, 409–416.

(36) Just, A. C.; Wright, R. O.; Schwartz, J.; Coull, B. A.; Baccarelli, A. A.; Tellez-Rojo, M. M.; Moody, E.; Wang, Y.; Lyapustin, A.; Kloog, I. Using High-Resolution Satellite Aerosol Optical Depth To

Estimate Daily PM2.5 Geographical Distribution in Mexico City. *Environ. Sci. Technol.* **2015**, *49*, 8576–8584.

(37) Lee, H. J.; Chatfield, R. B.; Strawa, A. W. Enhancing the applicability of satellite remote sensing for PM2.5 estimation using MODIS deep blue AOD and land use regression in california, united states. *Environ. Sci. Technol.* **2016**, *50*, 6546–6555.

(38) Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; de Hoogh, K.; De Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M.; Scortichini, M. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013-2015, using a spatiotemporal land-use random-forest model. *Environ. Int.* **2019**, *124*, 170–179.

(39) Shtein, A.; Karnieli, A.; Katra, I.; Raz, R.; Levy, I.; Lyapustin, A.; Dorman, M.; Broday, D. M.; Kloog, I. Estimating daily and intra-daily PM10 and PM2.5 in Israel using a spatio-temporal hybrid modeling approach. *Atmos. Environ.* **2018**, *191*, 142–152.

(40) Shtein, A.; Kloog, I.; Schwartz, J.; Silibello, C.; Michelozzi, P.; Gariazzo, C.; Viegi, G.; Forastiere, F.; Karnieli, A.; Just, A. C.; Stafoggia, M. Estimating Daily PM2.5 and PM10 over Italy Using an Ensemble Model. *Environ. Sci. Technol.* **2019**, *54*, 120–128.

(41) Zheng, C.; Zhao, C.; Zhu, Y.; Wang, Y.; Shi, X.; Wu, X.; Chen, T.; Wu, F.; Qiu, Y. Analysis of influential factors for the relationship between $PM_{2.5}$ and AOD in Beijing. *Atmos. Chem. Phys.* **2017**, *17*, 13473–13489.

(42) Stirnberg, R.; Cermak, J.; Andersen, H. An Analysis of Factors Influencing the Relationship between Satellite-Derived AOD and Ground-Level PM10. *Remote Sens (Basel)* **2018**, *10*, 1353.

(43) Zhang, X.; Chu, Y.; Wang, Y.; Zhang, K. Predicting daily PM2.5 concentrations in Texas using high-resolution satellite aerosol optical depth. *Sci. Total Environ.* **2018**, *631-632*, 904–911.

(44) Xie, Y.; Wang, Y.; Bilal, M.; Dong, W. Mapping daily PM2.5 at 500 m resolution over Beijing with improved hazy day performance. *Science of The Total Environment* **2019**, *659*, 410–418.

(45) Chatfield, R. B.; Sorek-Hamer, M.; Esswein, R. F.; Lyapustin, A. Satellite mapping of $PM_{2.5}$ episodes in the wintertime San Joaquin Valley: a “static” model using column water vapor. *Atmos. Chem. Phys.* **2020**, *20*, 4379–4397.

(46) Song, W.; Jia, H.; Huang, J.; Zhang, Y. A satellite-based geographically weighted regression model for regional PM2.5 estimation over the Pearl River Delta region in China. *Remote Sens. Environ.* **2014**, *154*, 1–7.

(47) Ma, Z.; Hu, X.; Huang, L.; Bi, J.; Liu, Y. Estimating ground-level PM2.5 in China using satellite remote sensing. *Environ. Sci. Technol.* **2014**, *48*, 7436–7444.

(48) Guo, Y.; Tang, Q.; Gong, D.-Y.; Zhang, Z. Estimating ground-level PM2.5 concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. *Remote Sens. Environ.* **2017**, *198*, 140–149.

(49) Sorek-Hamer, M.; Strawa, A. W.; Chatfield, R. B.; Esswein, R.; Cohen, A.; Broday, D. M. Improved retrieval of PM2.5 from satellite data products using non-linear methods. *Environ. Pollut.* **2013**, *182*, 417–423.

(50) You, W.; Zang, Z.; Zhang, L.; Zhang, M.; Pan, X.; Li, Y. A nonlinear model for estimating ground-level PM10 concentration in Xi'an using MODIS aerosol optical depth retrieval. *Atmos. Res.* **2016**, *168*, 169–179.

(51) Li, L.; Chen, B.; Zhang, Y.; Zhao, Y.; Xian, Y.; Xu, G.; Zhang, H.; Guo, L. Retrieval of daily PM2.5 concentrations using nonlinear methods: A case study of the beijing–tianjin–hebei region, china. *Remote Sens (Basel)* **2018**, *10*, 2006.

(52) He, Q.; Huang, B. Satellite-based mapping of daily high-resolution ground PM 2.5' ' in China via space-time regression modeling. *Remote Sens. Environ.* **2018**, *206*, 72–83.

(53) Wei, J.; Huang, W.; Li, Z.; Xue, W.; Peng, Y.; Sun, L.; Cribb, M. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Remote Sens. Environ.* **2019**, *231*, 111221.

(54) Chen, M.-J.; Yang, P.-H.; Hsieh, M.-T.; Yeh, C.-H.; Huang, C.-H.; Yang, C.-M.; Lin, G.-M. Machine learning to relate PM2.5 and PM10 concentrations to outpatient visits for upper respiratory tract infections in Taiwan: A nationwide analysis. *World J Clin Cases* **2018**, *6*, 200–206.

(55) Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L. A.; Strickland, M. J.; Liu, Y. Estimating $PM_{2.5}$ Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ. Sci. Technol.* **2017**, *51*, 6936–6944.

(56) Lary, D. J.; Lary, T.; Sattler, B. Using machine learning to estimate global PM2.5 for environmental health studies. *Environ. Health Insights* **2015**, *9*, 41–52.

(57) Matsuki, K.; Kuperman, V.; Van Dyke, J. A. The Random Forests statistical technique: An examination of its value for the study of reading. *Sci. Stud. Read.* **2016**, *20*, 20–33.

(58) Chowdhury, S.; Dey, S.; Di Girolamo, L.; Smith, K. R.; Pillarisetti, A.; Lyapustin, A. Tracking ambient PM2.5 build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset. *Atmos. Environ.* **2019**, *204*, 142–150.

(59) Mandal, S.; Madhipatla, K. K.; Guttikunda, S.; Kloog, I.; Prabhakaran, D.; Schwartz, J. D. Ensemble averaging based assessment of spatiotemporal variations in ambient PM2.5 concentrations over Delhi, India, during 2010–2016. *Atmos. Environ.* **2020**, *224*, 117309.

(60) Mhawish, A.; Banerjee, T.; Broday, D. M.; Misra, A.; Tripathi, S. N. Evaluation of MODIS Collection 6 aerosol retrieval algorithms over Indo-Gangetic Plain: Implications of aerosols types and mass loading. *Remote Sens. Environ.* **2017**, *201*, 297–313.

(61) Kumar, M.; Parmar, K. S.; Kumar, D. B.; Mhawish, A.; Broday, D. M.; Mall, R. K.; Banerjee, T. Long-term aerosol climatology over Indo-Gangetic Plain: Trend, prediction and potential source fields. *Atmos. Environ.* **2018**, *180*, 37–50.

(62) Vinjamuri, K. S.; Mhawish, A.; Banerjee, T.; Sorek-Hamer, M.; Broday, D. M.; Mall, R. K.; Latif, M. T. Vertical distribution of smoke aerosols over upper Indo-Gangetic Plain. *Environ. Pollut.* **2020**, *257*, 113377.

(63) Levy, R. C.; Mattoo, S.; Munchak, L. A.; Remer, L. A.; Sayer, A. M.; Patadia, F.; Hsu, N. C. The Collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech.* **2013**, *6*, 2989–3034.

(64) Hsu, N. C.; Jeong, M. J.; Bettenhausen, C.; Sayer, A. M.; Hansell, R.; Seftor, C. S.; Huang, J.; Tsay, S. C. Enhanced Deep Blue aerosol retrieval algorithm: The second generation. *J. Geophys. Res. Atmos.* **2013**, *118*, 9296–9315.

(65) Yuval; Levi, Y.; Dayan, U.; Levy, I.; Broday, D. M. On the association between characteristics of the atmospheric boundary layer and air pollution concentrations. *Atmos. Res.* **2020**, *231*, 104675.

(66) Shafran-Nathan, R.; Yuval; Broday, D. M. Impacts of personal mobility and diurnal concentration variability on exposure misclassification to ambient pollutants. *Environ. Sci. Technol.* **2018**, *52*, 3520–3526.

(67) The international centre for tropical agriculture (CIAT). *PANS Pest Articles & News Summaries* **1971**, *17*, 277–279.

(68) Lee, H. J.; Liu, Y.; Coull, B. A.; Schwartz, J.; Koutrakis, P. A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations. *Atmos. Chem. Phys.* **2011**, *11*, 7991–8002.

(69) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2/3*, 18–22.

(70) Breiman, L. Random Forests. *Springer Science and Business Media LLC* **2001**.

(71) Nicodemus, K. K.; Malley, J. D.; Strobl, C.; Ziegler, A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* **2010**, *11*, 110.

(72) Rodríguez, J. D.; Pérez, A.; Lozano, J. A. Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575.

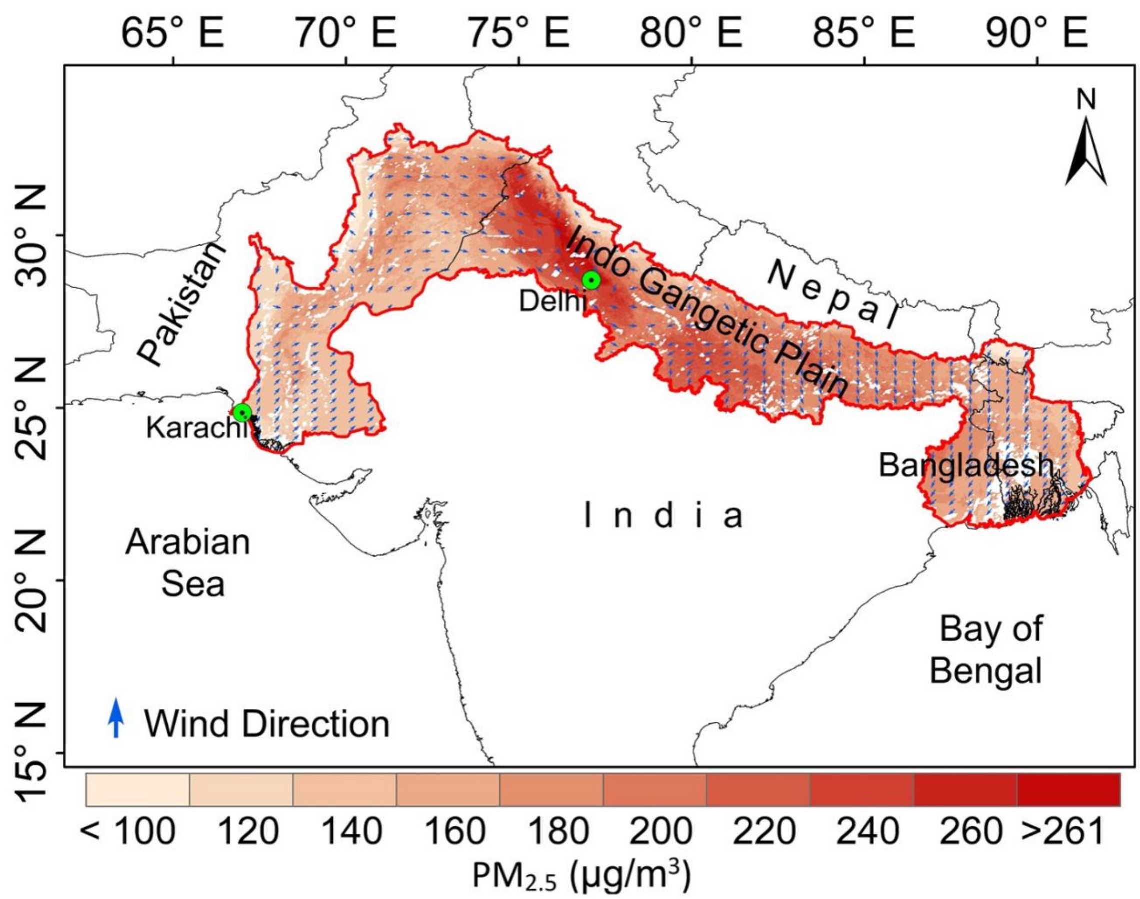
(73) Wang, J.; Martin, S. T. Satellite characterization of urban aerosols: Importance of including hygroscopicity and mixing state in the retrieval algorithms. *J. Geophys. Res.* **2007**, D17203, *112*.

(74) Xie, Y.; Wang, Y.; Zhang, K.; Dong, W.; Lv, B.; Bai, Y. Daily Estimation of Ground-Level PM2.5 Concentrations over Beijing Using 3 km Resolution MODIS AOD. *Environ. Sci. Technol.* **2015**, *49*, 12280–12288.

(75) Jethva, H.; Torres, O.; Field, R. D.; Lyapustin, A.; Gautam, R.; Kayetha, V. Connecting Crop Productivity, Residue Fires, and Air Quality over Northern India. *Sci. Rep.* **2019**, *9*, 16594.

(76) Singh, N., Banerjee, T., Raju, M.P., Deboudt, K., Sorek-Hamer, M., Singh, R.S. and Mall, R.K., 2018. Aerosol chemistry, transport, and climatic implications during extreme biomass burning emissions over the Indo-Gangetic Plain. Atmospheric Chemistry and Physics, 18(19), pp.14197-14215.

TOC graphic


65° E
70° E
75° E
80° E
85° E
90° E
30° N
25° N
20° N
15° N
N
Pakistan
Delhi
Indo Gangetic Plain
Nepal
Karachi
Bangladesh
India
Arabian Sea
Bay of Bengal
Wind Direction
< 100
120
140
160
180
200
220
240
260
>261
PM2.5 (μg/m3)