Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.210105

Machine Learning Algorithm for Estimating Surface $PM_{2.5}$ in Thailand

Pawan Gupta[1,2*], Shanshan Zhan[3], Vikalp Mishra[3,7], Aekkapol Aekakkararungroj[4], Amanda Markert[3,7], Sarawut Paibong[5], Farrukh Chishtie[4,6]

[1] Universities Space Research Association (USRA), Huntsville, USA
[2] Marshall Space Flight Center, Huntsville, AL, USA
[3] Earth System Science Center, The University of Alabama in Huntsville, Huntsville, AL, USA
[4] Asian Disaster Preparedness Center, Bangkok, Thailand
[5] Thai Pollution Control Department, Bangkok, Thailand
[6] Spatial Informatics Group, Pleasanton, CA, USA
[7] SERVIR Science Coordination Office, NASA Marshall Space Flight Center, Huntsville, AL, USA

ABSTRACT

We have used NASA's Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA2) reanalysis data of aerosols and meteorology into a machine learning algorithm (MLA) to estimate surface $PM_{2.5}$ concentration in Thailand. One year of hourly data from 51 ground monitoring stations in Thailand was spatiotemporally collocated with MERRA2 fields. The integrated data then used to train and validate a supervised MLA' random forest' to estimate hourly and daily $PM_{2.5}$ concentrations. The MLA is cross-validated using a 10-fold random sampling approach. The trained MLA can estimate $PM_{2.5}$ with close to zero mean bias across the country. The correlation coefficient of 0.95 with slope and intercept values of 0.95 and 0.88 are achieved between observed and estimated $PM_{2.5}$. The MLA also shows underestimation at hourly scale under very clean conditions ($PM_{2.5} < 10$ μg m^{-3}) and overestimation during high loading ($PM_{2.5} > 80$ μg m^{-3}). The hourly data also demonstrate high skill in following the diurnal cycle during different seasons of the year. The daily mean $PM_{2.5}$ (24-hour) values follow day-to-day variability very well (correlation coefficient of 0.98, RMSE = 3.14 μg m^{-3}), showing high value during winter months (November–February) and lower during other seasons. The trained MLA has the potential to reprocess the MERRA2 timeseries for the region, and the bias corrected data can be used in other applications such as long-term trend analysis and health exposure studies. The MLA can also be applied to GEOS forecasted fields to generate bias corrected air quality forecasts for the region.

Keywords: Thailand, MERRA2, $PM_{2.5}$, Air quality, Machine learning

OPEN ACCESS

Received: May 7, 2021
Revised: August 15, 2021
Accepted: September 10, 2021

* **Corresponding Author:**
pawan.gupta@nasa.gov

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Air quality is deteriorating in many urban and rural areas across the globe (*State of Global Air*, 2020). A recent World Bank study found that the effects of air pollution on health cost the world economy over $5 trillion in 2013 (World Bank, 2016). Air pollution ranks as the fourth most significant risk factor for fatalities worldwide behind high blood pressure, diet, and smoking (Brauer *et al.*, 2016). Fine particulate matter (PM or $PM_{2.5}$) has been linked to severe health problems, including heart and lung disease (Health Effects Institute, 2004; Mehta *et al.*, 2021; Pope III *et al.*, 2009, 2002; Samet *et al.*, 2000a, b, c; van Donkelaar *et al.*, 2015; World Health Organization, 2014) The $PM_{2.5}$ can be directly emitted in the atmosphere from fossil fuel burning, fires, dust and can also formed in the atmosphere by gas-particle phase chemistry. In Thailand, the primary sources of $PM_{2.5}$ in the region include seasonal agricultural and forest fires, coal-based power plants, industries, constructions, and transportation (Chirasophon and Pochanart, 2020;

Narita *et al.*, 2019; Uttamang *et al.*, 2018) and responsible for hazy and invisible skyline in cities, including Bangkok.

Traditionally, $PM_{2.5}$ is measured using ground monitors comprised of regulatory grade instruments, but their spatial coverage is often limited due to associated high costs. The next generation of air quality monitoring technologies, such as low-cost sensor networks and state-of-the-art satellite measurements, provide an alternative to the traditional and expensive surface monitors. However, at present, they have limited applications for regulatory purposes due to inadequate data availability and quality. In the past two decades, NASA and other space agencies around the world have relied on low earth orbiting (LEO) sensors (i.e., MODIS, MISR, OMI, VIIRS, CALIPSO) to obtain air quality related information, which is limited to at best one measurement per day per sensor during daylight, and cloud cover further limits the data availability. In more recent years, with the launch of advanced geostationary (GEO) satellites (i.e., GOES-R series, Himawari-08/09), continuous (minutes to an hour) air quality monitoring at high spatial and temporal resolution from space-based measurements are becoming a reality. However, GEO sensors are relatively few and only cover limited geographical locations.

Many research efforts have leveraged satellite-derived aerosol properties to estimate $PM_{2.5}$ at various spatial and temporal scales over the last two decades (Hoff and Christopher, 2009; Lee, 2020). Almost all of these efforts use satellite-derived Aerosol Optical Depth (AOD) in a statistical model (Gupta *et al.*, 2007; Gupta and Christopher, 2009a; Zhang and Kondragunta, 2021) or combine AOD with a chemical transport model (Liu *et al.*, 2005; van Donkelaar *et al.*, 2010, 2015; Ghude *et al.*, 2020) to estimate surface $PM_{2.5}$. Research on this topic has increased our understanding of how aerosol retrievals from the satellite can be effectively used to estimate surface $PM_{2.5}$ with known sources of uncertainties (i.e., local meteorology, aerosol retrievals). We also understand that in addition to AOD, meteorology, land use type, elevation, population, and other parameters in the statistical model can significantly improve the accuracies of $PM_{2.5}$ estimates (Lee, 2020). The AOD-$PM_{2.5}$ correlations are also affected by inherent uncertainties in AOD retrievals. Depending on the choice of AOD retrieval algorithm (i.e., Dark Target, Deep Blue, MAIAC, NOAA Enterprise), various factors play a role in AOD accuracies such as topography, seasonally changing surface characterization (i.e., vegetation cover, surface reflectance), aerosol type, size distribution and vertical distribution of aerosols in the atmospheric column along with sensor and solar viewing geometry.

Therefore, to overcome the limitations of using AOD alone to estimate $PM_{2.5}$, we implemented a combination of physical and statistical modeling with a machine learning algorithm (MLA) in Thailand. This approach used NASA's MERRA2 reanalysis (physical model) datasets on aerosols and meteorological fields to estimate surface $PM_{2.5}$ using an MLA (statistical model). In this approach, we rely on the MLA capabilities to define the complex non-linear relationships between surface $PM_{2.5}$, AOD, aerosol components, and the meteorological processes that directly or indirectly control the $PM_{2.5}$ concentration for the local condition at the surface.

2 DATA AND METHOD

2.1 Surface $PM_{2.5}$ Data

Particulate matter mass concentration (in $\mu g\ m^{-3}$) with aerodynamic diameters less than 2.5 μm ($PM_{2.5}$) is regularly monitored by the Thai Pollution Control Department (PCD) using an automated continuous system. Generally, according to the equivalent method from U.S. EPA Federal Reference Method (FRM), a Tapered Element Oscillating Microbalance (TEOM) instrument or Beta Attenuation Mass (BAM) monitor is used to measure the mass of $PM_{2.5}$ particles in units of $\mu g\ m^{-3}$. These instruments measure the mass concentration of particulate matter near the surface and considered gold standard providing ground truth. $PM_{2.5}$ data from these networks include 24 h average (daily) concentration data and continuous (hourly) $PM_{2.5}$ mass concentration measurements. Based on the data availability at the time of study, we used $PM_{2.5}$ data for year 2018 from 51 active $PM_{2.5}$ stations located in Thailand (Fig. 4) to train and validate a machine-learning algorithm.

2.2 MERRA2 Reanalysis

The Goddard Earth Observing System Model (GEOS) data assimilation system operated by

NASA's Global Modeling and Assimilation Office (GMAO) creates meteorological and aerosol analyses and forecasts in real-time (https://gmao.gsfc.nasa.gov). The aerosol products are derived from an online version of the Goddard Chemistry Aerosol Radiation and Transport (GOCART) aerosol model that includes organic carbon, black carbon, sulfate, dust, and sea spray aerosols (Chin *et al.*, 2002; Colarco *et al.*, 2010). More recently, nitrate has also been added to the aerosol component list in the forward processing and brown carbon is also being considered to account for biomass burning (Hammer *et al.*, 2016; Buchard *et al.*, 2017). The fire emissions used in GEOS are from the fire radiative power (FRP) based Quick Fire Emissions Dataset (QFED) (Darmenov and da Silva, 2015). It uses an Earth System Model Framework (ESMF) with aerosols and chemistry coupled to a Global Climate Model (GCM). It has consistent processing of earth system observations using a model, unchanged data assimilation system; and it assimilates millions of bias-corrected aerosol measurements from multiple satellites (MODIS, MISR, AVHRR) and surface monitoring stations (e.g., AERONET) along with meteorology. GEOS does not explicitly report $PM_{2.5}$ mass concentration near the surface. However, from adding the mass concentrations of different aerosol components for the surface layer, we can derive $PM_{2.5}$. Such $PM_{2.5}$ data derived from GEOS system and processed as reanalysis has been validated (Buchard *et al.*, 2017, 2016; Provençal *et al.*, 2017a, b) over the United States, Europe, Israel, and Taiwan, and it was found that MODIS AOD assimilation leads to better simulated surface $PM_{2.5}$ data compared to a similar version of the model with no aerosol data assimilation. The $PM_{2.5}$ values compare well when averaged over a larger area and longer time scales but can have large biases for diurnal, day-to-day, and seasonal averaging (Buchard *et al.*, 2016). This study used the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA2). MERRA2 provides data beginning in 1980 until the current time (Buchard *et al.*, 2017; Randles *et al.*, 2017). We used meteorological parameters (tavg1_2d_slv_Nx) and aerosol components (tavg1_2d_aer_Nx) from MERRA2 at the model output resolution of $0.5^\circ \times 0.625^\circ$ latitude by longitude. More details on specific parameters are provided in Section 2.4, and details on MERRA2 reanalysis can be found elsewhere (Buchard *et al.*, 2017; Randles *et al.*, 2017).

2.3 Spatiotemporal Collocation

The MERRA2 aerosols components, and meteorological parameters are available for every hour with a coarser spatial resolution, whereas the ground $PM_{2.5}$ data are point measurements with one hour frequency. Therefore, we integrated the three data sets into a single harmonized dataset using spatiotemporal collocation followed by the method reported in our earlier work (Gupta *et al.*, 2018; Gupta and Christopher, 2009b). In this method, we obtained ground $PM_{2.5}$ measurement corresponding to the nearest hour of MERRA2 output time to match two data sets temporally for every hour and day of the year. Similarly, we choose the MERRA2 grid cell nearest to the ground location by calculating the spherical distance (distance between two points on the surface of Earth) between the ground station and the center of the MERRA2 grid. It is important to note that MERRA2 data represents an average value over a larger area and is defined by its resolution, whereas ground monitoring is a point measurement. Also, due to the coarse resolution of MERRA2, it is possible that in areas with a high density of ground monitors, the same grid value is assigned to multiple ground locations. Our analysis is limited at model grid resolution and sub-grid variability in geophysical parameters are assumed uniform in this study.

2.4 Parameter Selection and MLA Development

Fig. 1 provides a flowchart of data integration, MLA training, and validation steps. In addition to AOD, other parameters (like temperature, pressure, relative humidity etc.) are used to account for additional spatiotemporal factors (e.g., local meteorology and emissions) that can influence $PM_{2.5}$ at the surface and thus the AOD-$PM_{2.5}$ relationships (Gupta and Christopher, 2006, 2009a, b; Marsha and Larkin, 2019). The meteorological parameters that strongly influence $PM_{2.5}$ include temperature, relative humidity, and height of the planetary boundary layer (Seinfeld and Pandis, 2006). Other processes that impact $PM_{2.5}$ concentration include small- to large-scale transport by winds, horizontal and vertical dispersion, and temperature gradients. The variations in available sunlight for photochemical reactions due to clouds and seasons, and available moisture also impact $PM_{2.5}$ concentration at the surface. And most importantly, the dilution of pollution in the

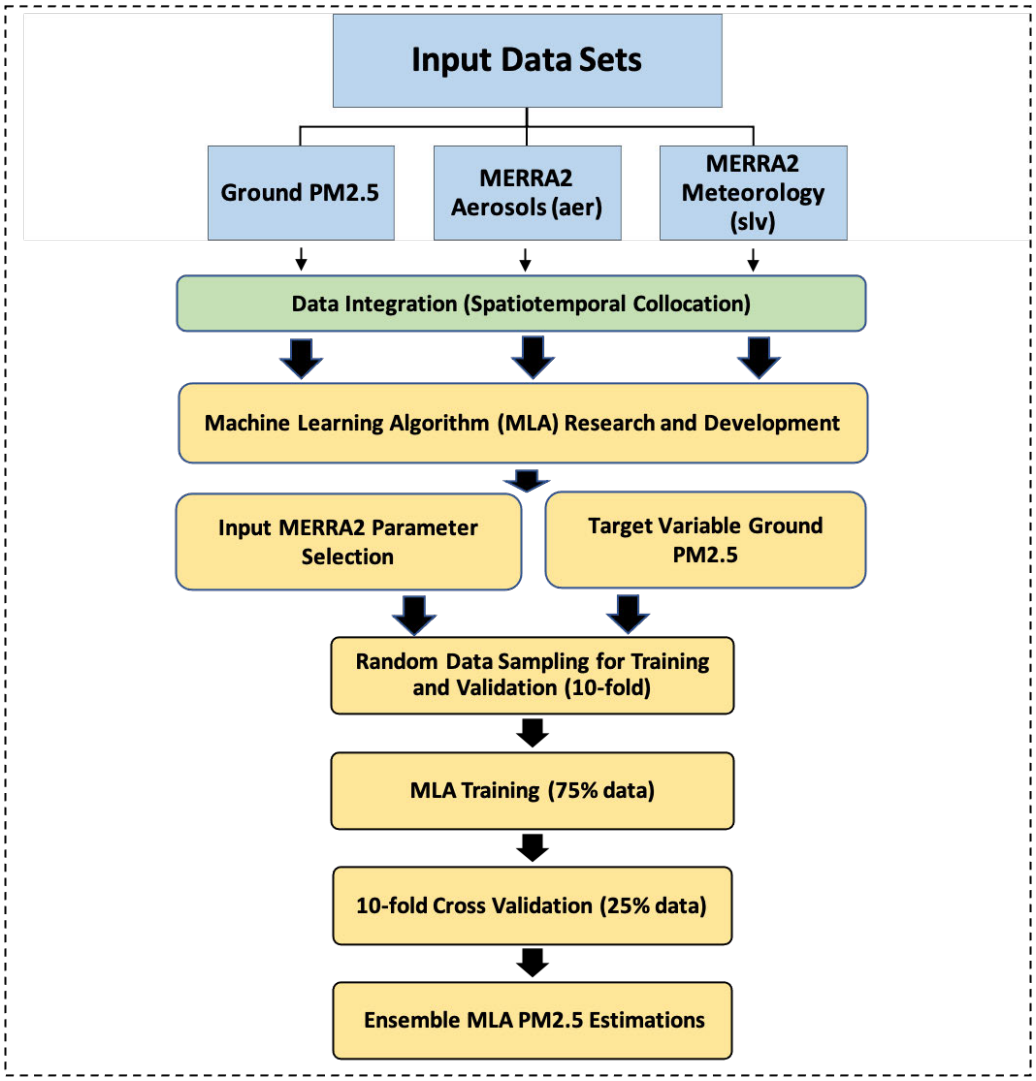

Fig. 1. Schematic of data integration and machine learning algorithm processes.

atmospheric boundary layer due to changes in vertical mixing. The variability in these meteorological conditions is primarily governed by large-scale high- and low-pressure systems, diurnal heating and cooling, and topography. Temperature can enhance the photochemical reactions in the atmosphere and hence the production of $PM_{2.5}$ particles (Seinfeld and Pandis, 2006). Temperature inversion can also reduce the vertical mixing and therefore increase the chemical concentration of precursors (Seinfeld and Pandis, 2006) gases. The higher concentration of precursors produces faster and more efficient chemical processes that convert gaseous emissions into particles. High relative humidity can enhance the growth and production of secondary particles and hence change the size distribution of particles and change the optical properties by modifying scattering efficiencies (Seinfeld and Pandis, 2006; Gupta and Christopher, 2009a; Wang and Martin, 2007; Zhang *et al*., 2021). Thus, meteorological parameters are included in the algorithm to account for atmospheric and surface conditions that may affect AOD and $PM_{2.5}$ differently. In addition, total AOD and mass concentrations of aerosols components are used as input to constrain the particles in the column and at near-surface levels. Fig. 2 provides the frequency distribution of input and output parameters. Here QV10m, Q500, and Q850 are specific humidity at 10 meters, 500 mb, and 850 mb pressure level, similarly T10m, T500, and T850 are air temperature for the same levels, and WIND is the wind speed at the surface. The aerosols' mass concentrations include Black Carbon (BCSMASS), Dust $PM_{2.5}$ (DUSMASS25), Organic Carbon (OCSMASS), Sulfur Dioxide (SO2SMASS), SO_4 mass (SO4SMASS), Sea Salt (SSSMASS25), and Total Extinction Aerosol Optical Depth at 550 nm (TOTEXTTAU). In addition, variations in geographical (i.e., latitude and longitude) and temporal

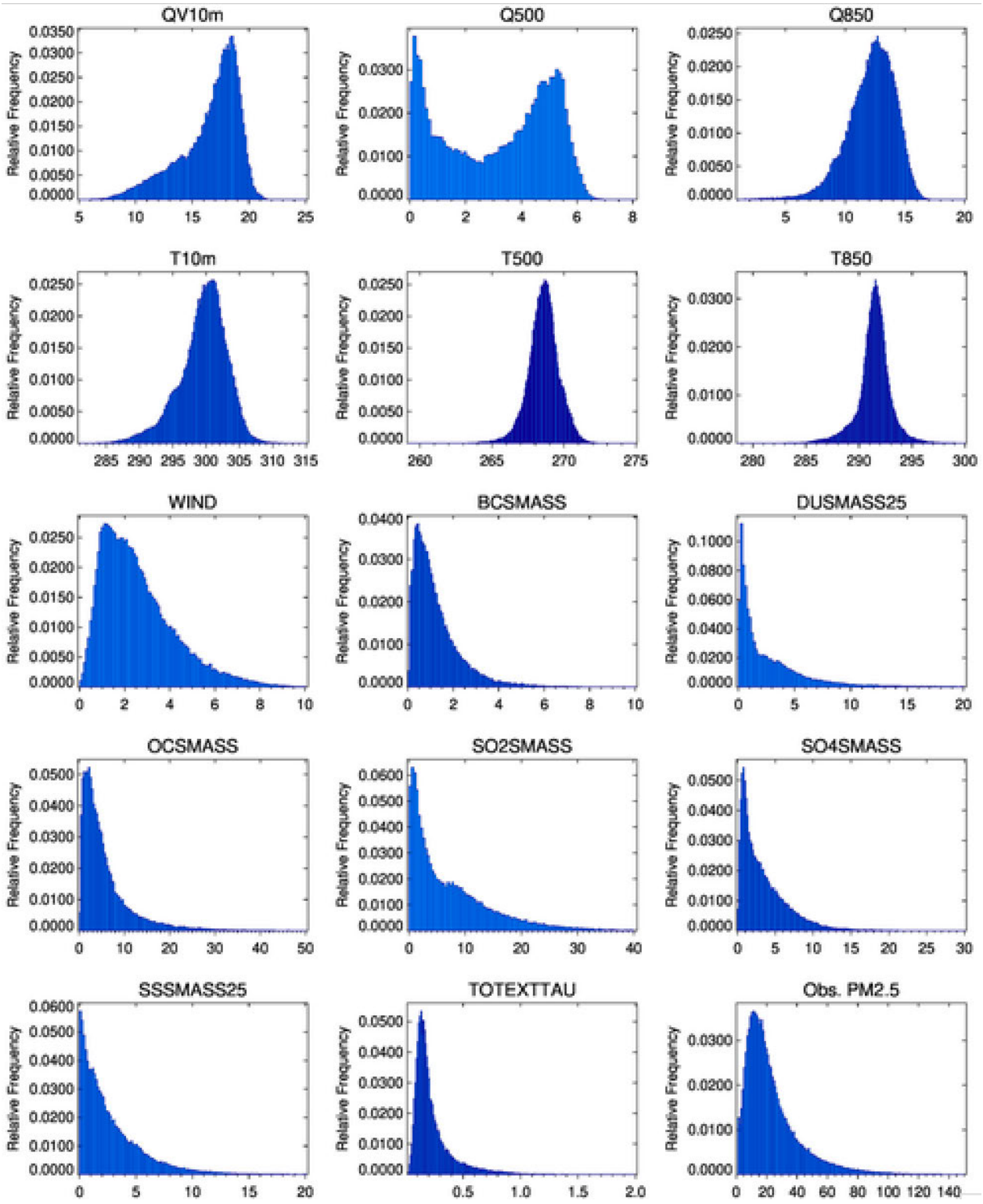

Fig. 2. Frequency distribution of input and output parameter for the machine learning algorithm.

(season, and time of the day) information have been used as input. The Obs. $PM_{2.5}$ is the output (or target) variable.

This integrated data set, which includes inputs and output, contains 266,094 samples, are used to train and validate the MLA. Several (regression, gradient boosting) ML algorithms have been tested and based on the performance; we selected random forest (RF) as a candidate ML algorithm for the detailed analysis in this study.

We used Scikit-learn (sklearn) machine learning library in python (https://scikit-learn.org). This free library provides various algorithms like k-neighbors, support vector machines, and RF. Before deciding on a particular algorithm, we tested XGBoost, Linear Regressor, SVM, and RF algorithm and based on the performance (not shown here), we selected RF as candidate algorithm for this study. The RF algorithm is a supervised MLA and one of the most used in modeling air quality using satellite remote sensing data sets (Masih, 2019 and references therein) due to its simplicity and diverse applications. It randomly samples a small subset from the dataset and uses it to train multiple decision trees using the bagging method. The bagging (or bootstrap aggregation) is ensemble learning technique and often used to control the noise in input datasets. The bagging method allows the combination of various learning methods which improve overall accuracy. The ensemble of decision trees (i.e., forest) is then used to produce the final output.

During the training, 266,094 data sample is divided randomly in 80% for training and 20% for validation and repeated training and validation by ten times. Figs. S1 and S2 provide scatter plots between observed and estimated $PM_{2.5}$ during this 10-fold training and validation exercises, respectively. The results (scatter and statistical parameters) indicate that the RF models perform consistent across the ten iterations and between training and validation. The consistency among 10-fold suggests the trained ML model attaining optimal performance and indicates less probability of overfitting or skewness in the outputs. The correlation between observed and estimated $PM_{2.5}$ across 10-fold training varies between 0.95 and 0.97, which is reduced to 0.88 to 0.92 for the validation data sets. The RMSE is about 4.7 to 5.6 $\mu g\ m^{-3}$ for training and jumped to 8.5 to 10.5 $\mu g\ m^{-3}$ in the validation data across 10-folds. An ensemble model is finally derived by averaging outputs from the 10 different models and used for further processing.

Next, we analyzed the relative importance of each input parameter on the $PM_{2.5}$ estimations (Fig. S3). The python tool used provides the measures of an input variable's importance by examining the tree node's role in reducing impurity across all trees in the forest. The relative importance score of each feature is computed and scaled automatically after training so that the sum of all importance is equal to one. The feature importance further provides guidelines on the use and impact of each input parameter to model target variable (i.e., $PM_{2.5}$).

2.5 Errors and Uncertainty Parameters

The number of data points (N), correlation coefficient (R), Root Mean Square Error (RMSE), Slope of the best-fit regression (M), Intercept (c), Mean Bias, and spatiotemporal consistency are among the few statistical parameters used to evaluate MLA performance throughout training and validation process. The R shows the goodness of fitting, while the Mean Bias symbolizes the differences in magnitude between estimated and ground truth. The RMSE represents the impact of extreme values and represents the variance of an MLA when comparing different MLA performances.

3 RESULTS AND DISCUSSION

The results presented here are from the ensemble of ten different machine learning models (MLM) trained to perform 10-fold cross-validation. Results and evaluation of the MLM were performed by statistical measures as discussed in Section 2.5. Time series of estimated and measured $PM_{2.5}$ were analyzed for accuracy assessment and intercomparisons with MERRA2 $PM_{2.5}$. The $PM_{2.5}$ mass from both training and validation data sets are shown as scatterplots. A separate analysis is performed for hourly and daily average $PM_{2.5}$ mass concentration since hourly data sets may not be available in many locations, and daily values are used to define air quality standards in many countries including in Thailand. The statistical performance parameters were analyzed over individual stations (Fig. 4), and diagnostic and prognostic errors as a function of observed and estimated $PM_{2.5}$ were quantified (Fig. 5). The seasonal diurnal cycle, evaluation of daily mean $PM_{2.5}$ and day-to-day variability in $PM_{2.5}$ for the study period are not discussed in detail in the main paper but provided as supplementary material (Figs. S4, S5, and S6).

3.1 Inter-comparison and Validation

Fig. 3 presents a density scatter plot between hourly observed and MERRA2 derived $PM_{2.5}$ (top

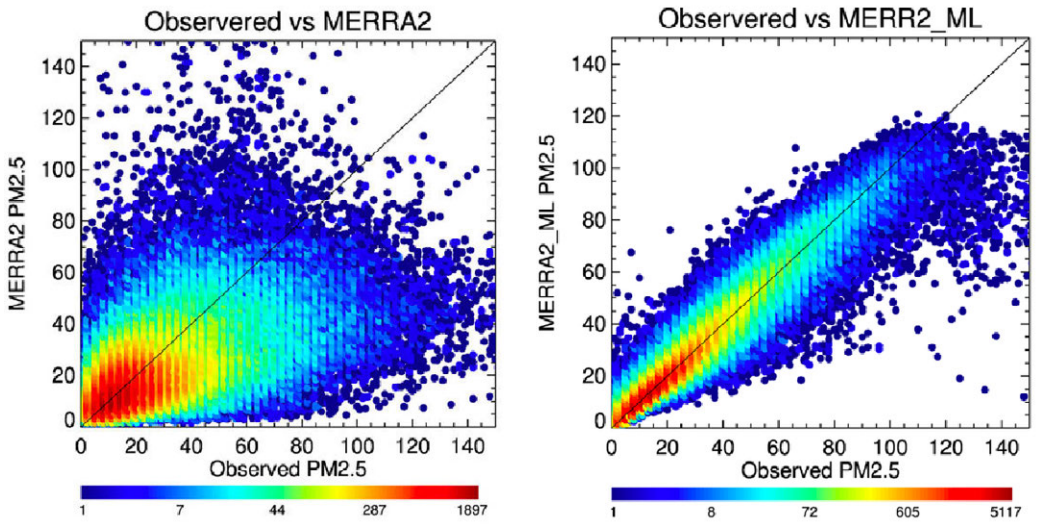

Fig. 3. The comparison between hourly observed (x-axis) and estimated $PM_{2.5}$ (y-axis). The top panel is for MERRA2 derived $PM_{2.5}$ whereas bottom panel is for MLA retrieved $PM_{2.5}$ using MERRA2 aerosols and meteorology as input. The statistical parameter for this analysis is presented in Table 1.

panel) and MERRA2_ML $PM_{2.5}$ (bottom panel). The MERRA2 $PM_{2.5}$ is calculated using the following equation (Malm *et al.*, 2011; Buchard *et al.*, 2016):

$$PM_{2.5} = DUST_{2.5} + SS_{2.5} + BC + 1.4 \times OC + 1.375 \times SO_4 \quad (1)$$

Here, DUST2.5, SS2.5, BC, OC, and SO_4 are respectively the surface concentrations of dust, sea-salt, back carbon, organic carbon, and sulfate particulates, all with diameter less or equal to 2.5 μm as reported in MERRA2 aerosol data. The data represents all collocated values (N = 266,094) from 51 different stations distributed across Thailand (Fig. 4). Table 1 reports the statistical parameter of the comparisons. It is apparent from the top panel that the MERRA2 $PM_{2.5}$ performs moderately against observed values with an overall correlation of 0.53 and RMSE of 16.7 μg m^{-3}. The linear fit's slope value of 0.38 indicates significant underestimation, specifically for observed $PM_{2.5}$ values larger than 30 μg m^{-3}. It is important to note that MERRA2 does not have nitrate and other tracee aerosol component in its $PM_{2.5}$ composition, leading to underestimation in areas with significant nitrates. The MERRA2_ML $PM_{2.5}$ compares excellently with the ground observations (bottom panel) except under a very high $PM_{2.5}$ concentration (> 100 μg m^{-3}). The correlation jumped to 0.95 with mean bias is close to zero (0.03 μg m^{-3}), and RMSE (5.9 μg m^{-3}) is reduced by about a factor of three as compared to MERRA2 $PM_{2.5}$ data. The slope value of 0.95 reflects minor underestimation by MLA as well. Overall, our MLA model performs better than the MERRA2 derived $PM_{2.5}$ given the accounting and inclusion of local observations including meteorological factors among others. We were surprised that the non-linearities of local interactions are better

Table 1. Statistical performance parameters for inter-comparison of MERRA2 and MERRA2_ML with observed $PM_{2.5}$ at hourly time scale. The scatter plot for these data are shown in Fig. 3.

Parameter	Data Source	
	MERRA2	MERRA2_ML
N	266094	266094
R	0.53	0.95
Bias	–3.8	0.03
RMSE	16.7	5.9
Slope (M)	0.38	0.95
Intercept (c)	9.0	0.88

represented in our MLA compared to global geophysical model, which also points to the importance of the locality of such factors for modeling surface $PM_{2.5}$ concentration levels.

3.2 Regional Distribution

The accuracy of the MERRA2_ML $PM_{2.5}$ estimations will be regionally and locally specific, depending on how well the input parameters such as meteorology and aerosol components match actual conditions. Local cloud conditions can limit the satellite retrievals and create a sampling bias within the MERRA2 data assimilation system and may introduce uncertainty into aerosols components. Furthermore, the spatiotemporal variability within the grid area may create biases in the collocation methodology that depends on the assumption of homogeneity in aerosols and meteorological fields. It is important to note that MERRA2 outputs grids are coarser (0.625 × 0.5 deg.) in spatial resolution, where ground monitors are point measurements. Here we evaluate the biases in derived MEERRA2_ML $PM_{2.5}$ over individual ground monitor locations.

We use only MEERRA2_ML $PM_{2.5}$ and calculate the same collocation statistics for each ground station individually for the regional and local analyses. Fig. 4 plots the values for correlation coefficient (R), mean bias (Bias), number of collocated samples (N), and RMSE for each station. It is clear from the figure that most ground stations are clustered in the south (i.e., around Bangkok city) and northern part of Thailand leaving significant monitoring gaps over rest of the country. The value of N varies from 1196 to 8706 in a different part of the country except for one station where we only had 272 collocated data points. The low number was associated with a lack of ground monitoring data availability during a certain period of the year due to technical and logistic reasons.

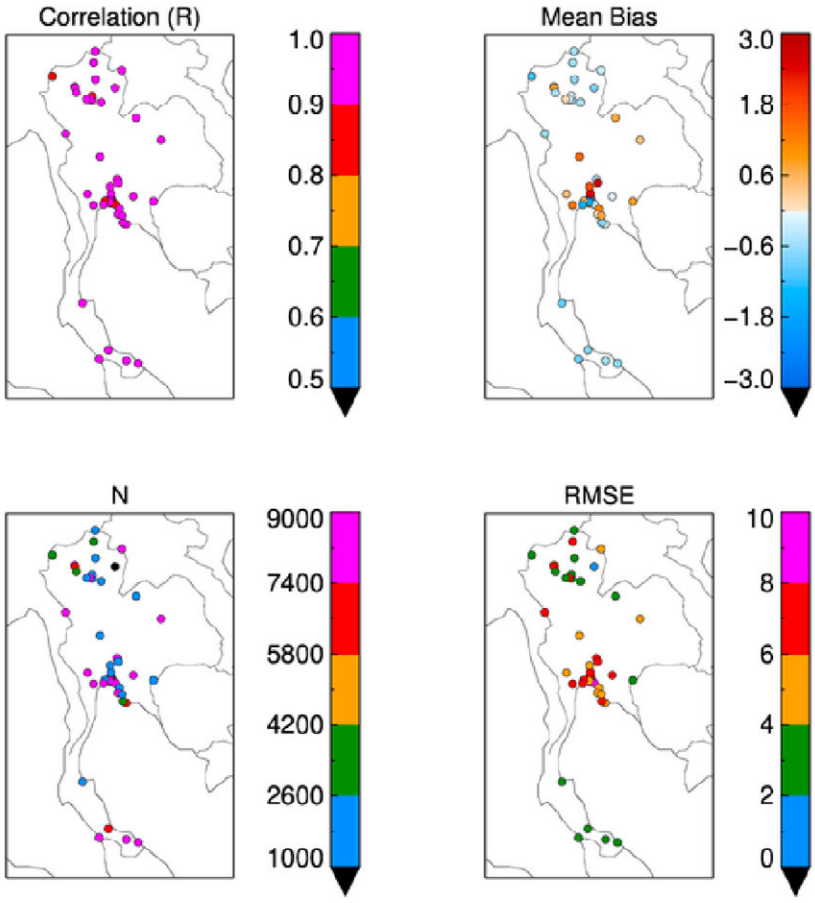

Fig. 4. The regional performance of MERRA2_ML $PM_{2.5}$ against observations at individual ground monitoring locations. The statistical performance parameters include: linear correlation coefficient (R, to left), mean bias (top right), number of collocated samples (N, bottom left), and root mean square error (RMSE, bottom right).

In general, the MERRA2_ML $PM_{2.5}$ shows high correlations ($0.87 < R < 0.98$) over much of Thailand. The correlation values are greater than 0.9 for 90% of the stations. Not all stations with strong correlations exhibit small mean biases. The mean bias values vary between ±2.9 μg m^{-3} on individual stations with almost half of the stations (45%) showing positive bias and the other half (55%) have negative biases. The negative biases are mainly observed in the northern part of the country, where seasonal biomass burning significantly enhanced the hourly and daily $PM_{2.5}$ values. The positive biases are mainly concentrated around Bangkok, where several ground monitors are clustered in smaller areas. In these conditions, due to coarser resolution of the MERRA2 grids, multiple stations can be associated with the same MERRA2 grid and can contribute towards uncertainty in the estimations. The RMSE (Fig. 4, bottom right) values over individual stations vary between 1.8 and 10.0 μg m^{-3} and represents the spread (standard deviation) of residuals (prediction error). The spatial variability in the performance of MERRA2_ML derived $PM_{2.5}$ is mainly associated with the density of ground stations within MERRA2 grids, seasonal biomass burning impact on local air quality, and variability in environmental conditions. The MLA can be further improved by including additional input parameters such as fire detection from satellite and land cover type information. Also, improved MERRA2 grid resolution will certainly improve heterogeneity in the ground measurements or by developing a separate model for different landcover types (agricultural areas vs. the urban area vs. the forested area etc.).

3.3 Diagnosis and Prognosis Errors

We next explored the relationship between bias (Estimated-Observed) $PM_{2.5}$ range for the integrated data set. At each collocated pair, the observed $PM_{2.5}$ is subtracted from the MERRA2_ML $PM_{2.5}$ so that a positive difference indicates overestimation. The data is then sorted according to an observed (diagnosis) and estimated (prognosis) $PM_{2.5}$ in the database. The integrated pairs are grouped into 266 bins, each containing 1000 pairs, the mean of each bin is increasing order. Thus, there are equal numbers of data pairs in each bin, but the bins are not equally spaced along the x-axis. Fig. 5 shows bias as a function of observed $PM_{2.5}$ (top panel), called diagnosis error whereas bias, as a function of MERRA2_ML $PM_{2.5}$ (bottom panel) called prognostic errors. The mean (red), median (green), and standard deviations (shaded area) of the bias are calculated for each bin and presented in the figure. The diagnosis error helps us understand the actual performance of the MLA at the ground locations but does not provide any way to correct it. On the other hand, the prognosis error can be used to correct output at locations where the ground monitor may or may not be available. The top panel clearly shows that MERRA2_ML are positively biased under very clean conditions ($PM_{2.5} < 10$ μg m^{-3}), which

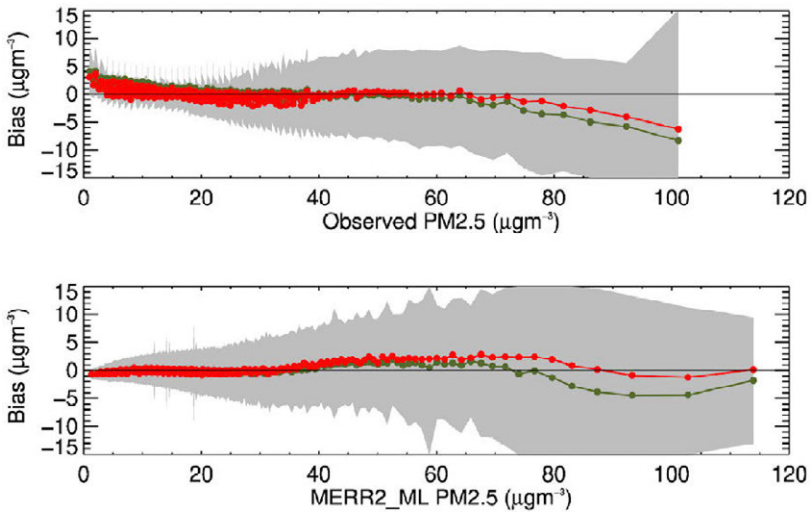

Fig. 5. The diagnostic (top pane, bias as function of observed $PM_{2.5}$) and prognostic (bottom, bias as function of MERRA2_ML $PM_{2.5}$) errors for hourly MERRA2_ML $PM_{2.5}$. The mean bias (y-axis) is calculated as (estimated – observed) $PM_{2.5}$. The red color shows bin average values whereas green shows median values.

was not very apparent in the analysis presented earlier in Section 3.1 and 3.2. The negative bias under high $PM_{2.5}$ loading (> 80 μg m^{-3}) is also more apparent and significant. The symmetric pattern between mean and median confirms the unskewed data distribution with fewer outliers in each bin. The negative biases under clean conditions are most likely associated with the uncertainties in MERRA2 aerosol components. MERRA2 assimilate satellite retrieved aerosol optical depth to constrain columnar aerosols in the atmosphere. It is well known that the satellite retrievals have difficulties in retrieving AODs under clean conditions and are often associated with higher uncertainties (Gupta *et al.*, 2016; Levy *et al.*, 2013). The underestimation for the high $PM_{2.5}$ values (> 80 μg m^{-3}) is associated with the under sampling of high $PM_{2.5}$ in the MLA training data set. The integrated data set only has about 1.8% instances with values larger than 80 μg m^{-3}, which is also apparent in the frequency distribution of observed $PM_{2.5}$ (Fig. 2). Most statistical methods, including MLA, try to learn the mathematical relationship between input and output parameters based on training data sets, and often, their performance skewed towards higher sampling density (Gupta and Christopher, 2009b; Lee, 2020). In this case, due to the low instances of high $PM_{2.5}$ values in the training data set, MLA produces better results for lower $PM_{2.5}$ values and underestimates at the high concentration range.

Finally, we also analyzed MERRA2_ML $PM_{2.5}$ at different temporal scales, including hourly diurnal cycle and 24-hour mean. When analyzed as a function of time of the day and day of the year, the errors do not show any dependency (not shown here). The diurnal pattern of observed MERRA2, MERRA2_ML were analyzed for each season (Fig. S4). The strongest diurnal cycle was found in winter (DJF) months, followed by post-monsoon (SON). The winter months have a higher $PM_{2.5}$ value during the nighttime and are most likely associated with low temperature and shallow boundary layer. The summer and spring months have very weak diurnal cycles. The MERRA2_ML $PM_{2.5}$ follows the diurnal cycle in each season to those in observed $PM_{2.5}$. Fig. S5 also shows the inter-comparison of daily mean observed and estimated $PM_{2.5}$ with an excellent correlation of 0.98 and slope value of 1.02. Fig. S6 demonstrates the capability of MERRA2_ML to capture day-to-day variability throughout the year and improvement compared with MERRA2. It is important to note that our analysis uses one year of data and therefore it is natural to ask how these MLM developed for 2018 perform for other years. In other words, does the MLM developed here, sufficiently produce consistent and accurate long-term timeseries over the region? To address these questions and produce a long-term timeseries of $PM_{2.5}$ datasets for the region, we have been analyzing data from 2010 to 2020 and expect to report results in future publications. Our preliminary analysis shows that MLA developed for 2018 is able to produce consistent performance over the years with no apparent year-to-year biases.

4 SUMMARY AND CONCLUSION

MERRA2 reanalysis by NASA's Global Modeling and Assimilation Office using the GEOS model is unique data set covering the global region and available from 1980 to the current time. The data sets include aerosols components and meteorological fields. Aerosols mass concentrations of individual components can be used to calculate a $PM_{2.5}$ equivalent mass concentration near the surface. The performance of MERRA2 calculated $PM_{2.5}$ around the world varies and depends on many factors, including missing emissions, the role of physical and chemical processes within the model, and its coarse resolution. We validate the MERRA2 $PM_{2.5}$ in Thailand using ground measurements and found that it significantly underestimates and fails to capture both diurnal and seasonal cycles.

Therefore, we used one year (2018) of hourly MERRA2 aerosols and meteorological parameters, $PM_{2.5}$ mass concentration from ground monitors, and performed spatiotemporal collocation to generate an integrated dataset. The integrated dataset was then used to train a machine learning algorithm to estimate surface $PM_{2.5}$ mass concentration at hourly and daily scales. The inputs to the algorithm are aerosols components, including aerosol optical depth and meteorological parameters, which directly or indirectly effects the $PM_{2.5}$ near the surface. The trained MLA is cross-validated using a 10-fold validation approach. The major finding of our research study are as follows:

- The MLA can produce hourly and daily mean $PM_{2.5}$ with very high accuracy consistently. The mean bias is close to zero, with correlation coefficients were higher than 0.9 in most cases.

- The MLA estimated $PM_{2.5}$ follows diurnal, day-to-day, and seasonal cycles as observed by ground monitors.
- The current MLA underestimates $PM_{2.5}$ under high $PM_{2.5}$ concentration (> 80 μg m^{-3}), limiting its application under very poor air quality conditions. The leading cause of underestimation by MLA is the lack of proper representation ($< 1.8\%$) of high values in the total data volume.
- We plan to further improve the MLA performance at higher range of $PM_{2.5}$ by including additional fire related satellite observations.
- The trained MLA can be applied to the long-term MERRA2 data sets, and spatiotemporal trends can be evaluated for the Thailand region.
- There is also potential to use trained MLA for bias correcting GEOS-FP $PM_{2.5}$ forecasts for the region.

ACKNOWLEDGMENTS

The authors would like to thank Thailand's Pollution Control Department for help accessing and using their ground-based $PM_{2.5}$ sensor network to support this study. The MERRA2 data were obtained from NASA's Earth Science Data Systems (https://earthdata.nasa.gov/). Pawan Gupta, and Shanshan Zhan were partially supported by the NASA ROSES program NNH17ZDA001N-TASNPP: The Science of Terra, Aqua, and Suomi NPP. The partial support for this work was provided by the joint US Agency for International Development (USAID) and National Aeronautics and Space Administration (NASA) initiative SERVIR-Mekong, Cooperative Agreement Number: AID-486-A-14-00002. Individuals affiliated with the University of Alabama in Huntsville (UAH) are funded through the NASA Applied Sciences Capacity Building Program, NASA Cooperative Agreement: NNM11AA01A. The author also like to thanks Mr. Abhishek Dutt of Trustnet Technologies for advising and providing critical inputs on use of machine learning algorithms.

Author Contributions: Conceptualization, PG; software, SZ, PG.; formal analysis, PG, SZ; validation, VM; writing—original draft preparation, PG.; writing—review and editing, VM, AA, AM, SP, FC; visualization, PG; supervision, PG; project administration, PG, AM, AA.; funding acquisition, PG, AM, AA. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.210105

REFERENCES

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R.V., Dentener, F., Dingenen, R. van, Estep, K., Amini, H., Apte, J.S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S., *et al.* (2016). Ambient air pollution exposure estimation for the global burden of disease 2013. Environ. Sci. Technol. 50, 79–88. https://doi.org/10.1021/acs.est.5b03709

Buchard, V., da Silva, A.M., Randles, C.A., Colarco, P., Ferrare, R., Hair, J., Hostetler, C., Tackett, J., Winker, D. (2016). Evaluation of the surface $PM_{2.5}$ in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. Atmos. Environ. 125, 100–111. https://doi.org/10.1016/j.atmosenv.2015.11.004

Buchard, V., Randles, C.A., da Silva, A.M., Darmenov, A., Colarco, P.R., Govindaraju, R., Ferrare, R., Hair, J., Beyersdorf, A.J., Ziemba, L.D., Yu, H. (2017). The MERRA-2 aerosol reanalysis, 1980 onward. Part II: Evaluation and case studies. J. Clim. 30, 6851–6872. https://doi.org/10.1175/JCLI-D-16-0613.1

Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B.N., Duncan, B.N., Martin, R.V., Logan, J.A., Higurashi, A., Nakajima, T. (2002). Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and sun photometer measurements. J. Atmos. Sci. 59, 461–483. https://doi.org/10.1175/1520-0469(2002)059<0461:TAOTFT>2.0.CO;2

Chirasophon, S., Pochanart, P. (2020). The long-term characteristics of PM_{10} and $PM_{2.5}$ in Bangkok, Thailand. Asian J. Atmos. Environ. 14, 73–83. https://doi.org/10.5572/ajae.2020.14.1.073

Darmenov, A.S., Da Silva, A.M. (2015). The Quick Fire Emissions Dataset (QFED): Documentation of versions 2.1, 2.2 and 2.4. NASA Technical Report Series on Global Modeling and Data Assimilation 38 (NASA/TM–2015–104606).

Ghude, S.D., Kumar, R., Jena, C., Debnath, S., Kulkarni, R.G., Alessandrini, S., Rajeevan, M. (2020). Evaluation of PM2.5 forecast using chemical data assimilation in the WRF-CHEM model: A novel initiative under the ministry of earth sciences air quality early warning system for Delhi, India. Curr. Sci. 118, 1803–1815. https://doi.org/10.18520/cs/v118/i11/1803-1815

Gupta, P., Christopher, S.A., Box, M.A., Box, G.P. (2007). Multi year satellite remote sensing of particulate matter air quality over Sydney, Australia. Int. J. Remote Sens. 28, 4483–4498. https://doi.org/10.1080/01431160701241738

Gupta, P., Christopher, S.A. (2009a). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. J. Geophys. Res. 114, D14205. https://doi.org/10.1029/2008JD011496

Gupta, P., Christopher, S.A. (2009b). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. J. Geophys. Res. 114, D20205. https://doi.org/10.1029/2008JD011497

Gupta, P., Levy, R.C., Mattoo, S., Remer, L.A., Munchak, L.A. (2016). A surface reflectance scheme for retrieving aerosol optical depth over urban surfaces in MODIS Dark Target retrieval algorithm. Atmos. Meas. Tech. 9, 3293–3308. https://doi.org/10.5194/amt-9-3293-2016

Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., Polidori, A., Kiros, F., Mills, K.C. (2018). Impact of California fires on local and regional air quality: The role of a low-cost sensor network and satellite observations. GeoHealth 2, 172–181. https://doi.org/10.1029/2018GH000136

Hammer, M.S., Martin, R.V., van Donkelaar, A., Buchard, V., Torres, O., Ridley, D.A., Spurr, R.J.D. (2016). Interpreting the ultraviolet aerosol index observed with the OMI satellite instrument to understand absorption by organic aerosols: Implications for atmospheric oxidation and direct radiative effects. Atmos. Chem. Phys. 16, 2507–2523. https://doi.org/10.5194/acp-16-2507-2016

Health Effects Institute (2004). Health effects of outdoor air pollution in developing countries of Asia: A literature review (No. 15). Health Effects Institue, Boston, MA, USA.

Hoff, R.M., Christopher, S.A. (2009). Remote sensing of particulate pollution from space: Have we reached the promised land? J. Air Waste Manage. Assoc. 59, 645–675. https://doi.org/10.3155/1047-3289.59.6.645

Lee, H.J. (2020). Advancing exposure assessment of $PM_{2.5}$ using satellite remote sensing: A review. Asian J. Atmos. Environ. 14, 319–334. https://doi.org/10.5572/ajae.2020.14.4.319

Levy, R.C., Mattoo, S., Munchak, L.A., Remer, L.A., Sayer, A.M., Patadia, F., Hsu, N.C. (2013). The Collection 6 MODIS aerosol products over land and ocean. Atmos. Meas. Tech. 6, 2989–3034. https://doi.org/10.5194/amt-6-2989-2013

Liu, Y., Sarnat, J.A., Kilaru, V., Jacob, D.J., Koutrakis, P. (2005). Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing. Environ. Sci. Technol. 39, 3269–3278. https://doi.org/10.1021/es049352m

Marsha, A., Larkin, N.K. (2019). A statistical model for predicting $PM_{2.5}$ for the western United States. J. Air Waste Manage. Assoc. 69, 1215–1229. https://doi.org/10.1080/10962247.2019.1640808

Masih, A. (2019). Machine learning algorithms in air quality modeling. Global J. Environ. Sci. Manage. 5, 515–534. https://doi.org/10.22034/GJESM.2019.04.10

Mehta, U., Dey, S., Chowdhury, S., Ghosh, S., Hart, J.E., Kurpad, A. (2021). The association between ambient $PM_{2.5}$ exposure and anemia outcomes among children under five years of Age in India. Environ. Epidemiol. 5, e125. https://doi.org/10.1097/EE9.0000000000000125

Narita, D., Oanh, N.T.K., Sato, K., Huo, M., Permadi, D.A., Chi, N.N.H., Ratanajaratroj, T., Pawarmart, I. (2019). Pollution characteristics and policy actions on fine particulate matter in a growing Asian economy: The case of Bangkok metropolitan region. Atmosphere 10, 227. https://doi.org/10.3390/atmos10050227

Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, K.I., Thurston, G.D. (2002). Lung

cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. J. Am. Med. Assoc. 287, 1132–1141. https://doi.org/10.1001/jama.287.9.1132

Pope III, C.A., Ezzati, M., Dockery, D.W. (2009). Fine-particulate air pollution and life expectancy in the United States. N. Engl. J. Med. 360, 376–386. https://doi.org/10.1056/NEJMsa0805646

Provençal, S., Buchard, V., da Silva, A.M., Leduc, R., Barrette, N. (2017a). Evaluation of PM surface concentrations simulated by Version 1 of NASA's MERRA Aerosol Reanalysis over Europe. Atmos. Pollut. Res. 8, 374–382. https://doi.org/10.1016/j.apr.2016.10.009

Provençal, S., Buchard, V., da Silva, A.M., Leduc, R., Barrette, N., Elhacham, E., Wang, S.H. (2017b). Evaluation of $PM_{2.5}$ surface concentrations simulated by version 1 of NASA's MERRA aerosol reanalysis over Israel and Taiwan. Aerosol Air Qual. Res. 17, 253–261. https://doi.org/10.4209/aaqr.2016.04.0145

Randles, C.A., da Silva, A., Buchard, V., Darmenov, A., Colarco, P.R., Aquila, V., Bian, H., Nowottnick, E.P., Pan, X., Smirnov, A., Yu, H., Govindaraju, R. (2017). The MERRA-2 Aerosol Assimilation (No. NASA/TM-2016-104606). National Aeronautics and Space Administration, USA.

Samet, J.M., Dominici, F., Curriero, F.C., Coursac, I., Zeger, S.L. (2000a). Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. N. Engl. J. Med. 343, 1742–1749. https://doi.org/10.1056/NEJM200012143432401

Samet, J.M., Dominici, F., Zeger, S.L., Schwartz, J., Dockery, D.W. (2000b). National Morbidity, Mortality, and Air Pollution Study. Part I: Methods and Methodologic Issues (No. 94–1). Health Effects Institute, Boston, MA, USA.

Samet, J.M., Zeger, S.L., Dominici, F., Curriero, F., Dockery, D.W., Schwartz, J., Zanobetti, A. (2000c). The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and mortality from air pollution in the United States (No. 94–2). Health Effects Institute, Boston, MA, USA.

Seinfeld, J.H., Pandis, S.N. (2006). Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 2nd ed. John Wiley & Sons, New York.

State of Global Air (2020). Health Effects Institute, Boston, MA, USA.

Uttamang, P., Aneja, V.P., Hanna, A.F (2018). Assessment of gaseous criteria pollutants in the Bangkok metropolitan region, Thailand. Atmos. Chem. Phys. 18, 12581–12593. https://doi.org/10.5194/acp-18-12581-2018

van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. Environ. Health Perspect. 118, 847–855. https://doi.org/10.1289/ehp.0901623

van Donkelaar, A., Martin, R.V., Brauer, M., Boys, B.L. (2015). Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environ. Health Perspect. 123, 135–143. https://doi.org/10.1289/ehp.1408646

Wang, J., Martin, S.T. (2007). Satellite characterization of urban aerosols: Importance of including hygroscopicity and mixing state in the retrieval algorithms. J. Geophys. Res. 112, D17203. https://doi.org/10.1029/2006JD008078

World Bank (2016). The Cost of Air Pollution: Strengthening the Economic Case for Action. World Bank Institute for Health Metrics and Evaluation, Washington, DC, USA.

World Health Organization (2014). World health statistics 2014. World Health Organization, Geneva. https://www.who.int/docs/default-source/gho-documents/world-health-statistic-reports/world-health-statistics-2014.pdf

Zhang, C., Ma, N., Fan, F., Yang, Y., Größ, J., Yan, J., Bu, L., Wang, Y., Wiedensohler, A. (2021). Hygroscopic growth of aerosol particles consisted of oxalic acid and its internal mixture with ammonium sulfate for the relative humidity ranging from 80% to 99.5%. Atmos. Environ. 252, 118318. https://doi.org/10.1016/j.atmosenv.2021.118318

Zhang, H., Kondragunta, S. (2021). Daily and hourly surface $PM_{2.5}$ estimation from satellite AOD. Earth Space Sci. 8, e2020EA001599. https://doi.org/10.1029/2020EA001599