Machine Learning Algorithm for Estimating Surface PM2.5 in Thailand

Pawan Gupta1,2*, Shanshan Zhan3, Vikalp Mishra3,7, Aekkapol Aekakkararungroj4, Amanda Markert3,7, Sarawut Paibong5, Farrukh Chishtie4,6

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

*Corresponding author: Tel: 1-2564687651
Email Address: pawan.gupta@nasa.gov

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 PM2.5 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 PM2.5 concentrations. The MLA is cross-validated using a 10-fold random sampling approach. The trained MLA can estimate PM2.5 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 PM2.5. The MLA also shows underestimation at hourly scale under very clean conditions (PM2.5<10 µgm⁻³) and overestimation during high loading (PM2.5>80 µgm⁻³). The hourly data also demonstrate high skill in following the diurnal cycle during different seasons of the year. The daily mean PM2.5 (24-hour) values follow day-to-day variability very well, showing high value during winter months (November to 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, PM2.5, Air Quality, Machine Learning.
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 PM2.5) 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; J. M. Samet et al., 2000, 2000a, 2000b; van Donkelaar et al., 2015; World Health Statistics, 2014) The PM2.5 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 PM2.5 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 primary pollutants responsible for hazy and invisible skyline in cities, including Bangkok.

Traditionally, PM2.5 are measured using ground monitors comprised of regulatory grade instruments, however, and 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 PM2.5 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) to estimate surface PM2.5. Research on this topic has increased our understanding of how aerosol retrievals from the satellite can be effectively used to estimate surface PM2.5 with known sources of uncertainties. 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 PM2.5 estimates (Lee, 2020). The AOD-PM2.5 correlations are also affected by inherent uncertainties in AOD retrievals. Depending on the choice of AOD retrieval algorithm (i.e., Dark Target, Deep Blue, MALAC, NOAA Enterprise), various factors play a role in AOD accuracies such as topography, seasonally changing surface characterization (i.e., NDVI, 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 PM25, 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 PM2.5 using an MLA (statistical model). In this approach, we rely on the MLA capabilities to define the complex non-linear relationships between
surface PM2.5, AOD, aerosol components, and the meteorological processes that directly or indirectly control the PM2.5 concentration for the local condition at the surface.

2 DATA AND METHOD

2.1 Surface PM2.5 data

Particulate matter mass concentration (in µgm⁻³) with aerodynamic diameters less than 2.5 µm (PM2.5) is regularly monitored by the Thai Pollution Control Department (PCD) using automated continuous system. Generally, according to the equivalent method from USEPA Federal Reference Method (FRM), a Tapered Element Oscillating Microbalance (TEOM) instrument or Beta Attenuation Mass (BAM) monitor is used to measure the mass of PM2.5 particles in units of µgm⁻³. These networks measure the mass concentration of gaseous and particulate air pollutants near the surface. PM2.5 data from these networks include 24 h average (daily) concentration data and continuous (hourly) PM2.5 mass concentration measurements. We used PM2.5 data for year 2018 from 51 active PM2.5 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, nitrate, dust, and sea spray aerosols, and will soon be extended to include brown carbon aerosols (Colarco et al., 2010). More recently, nitrate has also been added to the aerosol component list. The fire emissions used in GEOS are from the fire radiative power (FRP) based Quick Fire Emissions Dataset (QFED)
It uses an Earth System Model Framework (ESMF) with aerosols and chemistry coupled to Global Climate Model (GCM). It has consistent processing of earth system observations using a model, unchanged data assimilation system; c) 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 PM2.5 mass concentration near the surface. However, from adding the mass concentrations of different aerosol components for the surface layer, we can derive PM2.5. Such PM2.5 data derived from GEOS system and processed as reanalysis has been validated (Buchard et al., 2017, 2016; Provençal et al., 2017, 2017) over the United States, Europe, Israel, and Taiwan, and it was found that MODIS AOD assimilation leads to better simulated surface PM2.5 data compared to a similar version of the model with no aerosol data assimilation. The PM2.5 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° × 0.625° latitude by longitude. More details on specific parameters are provided in section xx, 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 PM2.5 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 et al., 2021, Thai Air

In this method, we obtained ground PM2.5 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 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.

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 PM2.5 at the surface and thus the AOD-PM2.5 relationships (Gupta and Christopher, 2006; 2009a 2009b; Marsha and Larkin, 2019). The meteorological parameters that strongly influence PM2.5 include temperature, relative humidity, and height of the planetary boundary layer (Seinfeld and Pandis, 2006). Other processes that impact PM2.5 concentration include small- to large-scale transport by winds, horizontal and vertical dispersion, variations in available sunlight for photochemical reactions due to clouds and seasons, temperature gradients, available moisture, and most importantly, the dilution of pollution in the 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 PM2.5 particles. Temperature inversion can also reduce the vertical mixing and therefore increase the chemical concentration of precursors. 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 (Gupta and Christopher, 2009a; Wang and Martin, 2007). Thus, meteorological parameters are included in the algorithm to account for atmospheric and surface conditions that may affect AOD and PM2.5 differently. Besides, 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 PM2.5 (DUSMASS25), Organic Carbon (OCSMASS), Sulfur Dioxide (SO2SMASS), SO4 mass (SO4SMASS), Sea Salt (SSSMASS25), Total Extinction Aerosol Optical Depth at 550 nm (TOTEXTTAU). In addition, variations in geographical and temporal information have been used as input. The Obs. PM2.5 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 Machine Learning Models (MLM). 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. 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 reference in it) due to its simplicity and diversity. It randomly samples a small subset from the dataset and uses this to train multiple decision trees
using the bagging method. 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, 266094 data sample is divided randomly in 80% for training and 20% for validation and repeated this step by ten times. Fig. S1 and S2 provide scatter plots between observed and estimated PM2.5 during this 10-fold training and validation exercises, respectively. The results (scatter and statistical parameters) indicate that the RF models' performance is consistent across the ten iterations and between training and validation. This suggests the trained model's attaining optimal performance and indicates less probability of overfitting or skewness in the outputs. The correlation between observed and estimated PM2.5 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 for training and jumped to 8.5 to 10.5 in the validation data across 10-folds. An ensemble model is finally chosen by averaging outputs from the 10 different models.

Next, we analyzed the relative importance of each input parameter on the PM2.5 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 provide guidelines on the use and impact of each input parameter to model target variable (i.e. PM2.5).

### 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.xx. Time series of estimated and measured PM2.5 were analyzed for accuracy assessment and intercomparisons with MERRA2 PM2.5. The PM2.5 mass from both training and validation data sets are shown as scatterplots. A separate analysis is performed for hourly and daily average PM2.5 mass concentration since hourly data sets may not be available in many locations, and daily values are used to define air quality standards. The statistical performance parameters were analyzed over individual stations (Fig. 4), and diagnostic and prognostic errors as a function of observed and estimated PM2.5 were quantified (Fig. 5). The seasonal diurnal cycle, evaluation of daily mean PM2.5 and day-to-day variability in PM2.5 for the study period are not discussed in details in the main paper but provided as supplementary material (Fig. S4, S5, and S6).

3.1 Inter-comparison and Validation

Fig 3 presents a density scatter plot between hourly observed and MERRA2 derived PM2.5 (top panel) and MERRA2_ML PM2.5 (bottom panel). The MERRA2 PM2.5 is calculated using the following equation (Malm et al., 2011, Buchard et al., 2016).

\[ PM2.5 = DUST_{2.5} + SS_{2.5} + BC + 1.4 \times OC + 1.375 \times SO_4 \]
Here, DUST2.5, SS2.5, BC, OC, and SO4 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=266094) 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 PM2.5 performs moderately against observed values with an overall correlation of 0.53 and RMSE of 16.7 μgm⁻³. The linear fit's slope value of 0.38 indicates significant underestimation, specifically for observed PM2.5 values larger than 30 μgm⁻³. It is important to note that MERRA2 does not have nitrate in its PM2.5 composition, leading to underestimation in areas with significant nitrates. The MERRA2_ML PM2.5 compares excellently with the ground observations (bottom panel) except under a very high PM2.5 concentration (>100 μgm⁻³). The correlation jumped to 0.95 with mean bias is close to zero (0.03 μgm⁻³), and RMSE (5.9 μgm⁻³) is reduced by about a factor of three as compared to MERRA2 PM2.5 data. The slope value of 0.95 reflects minor underestimation by MLA as well. Overall, our MLA model performs better than the MERRA2 derived PM2.5 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 represented in our MLA which also points to the importance of the locality of such factors for modeling surface PM2.5 concentration levels.

3.2 Regional Distribution

The accuracy of the MERRA2_ML PM2.5 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 (0.625x0.5 deg.) in size, where ground monitors are point measurements. Here we evaluate the biases in derived MEERRA2_ML PM2.5 over individual ground monitor locations.

We use only MEERRA2_ML PM2.5 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.

In general, the MERRA2_ML PM2.5 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 µgm⁻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 PM2.5 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 µgm⁻³ and represents the spread (standard deviation) of residuals (prediction error). The spatial variability in the performance of MERRA2_ML derived PM2.5 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 metros vs the forest etc.).

### 3.3 Diagnosis and Prognosis Errors

We next explored the relationship between bias (Estimated-Observed) PM2.5 range for the integrated data set. At each collocated pair, the observed PM2.5 is subtracted from the MERRA2_ML PM2.5 so that a positive difference indicates overestimation. The data is then sorted according to an observed (diagnosis) and estimated (prognosis) PM2.5 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 PM2.5 (top panel), called diagnosis error whereas bias, as a function of MERRA2_ML PM2.5 (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 (PM2.5<10), which was not very apparent in the analysis presented earlier in section 3.1 and 3.2. The negative bias under high PM2.5 loading (>80) 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 PM2.5 values (>80) is associated with the under sampling of high PM2.5 in the MLA training data set. The integrated data set only has about 1.8% instances with values larger than 80 µgm⁻³, which is also apparent in the frequency distribution of observed PM2.5 (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 PM2.5 values in the training data set, MLA produces better results for lower PM2.5 values and underestimates at the high concentration range.

Finally, we also analyzed MERRA2_ML PM2.5 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 PM2.5 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 PM2.5 follows the diurnal cycle in each season to those in observed PM2.5. Fig. S5 also shows the inter-comparison of daily mean observed and estimated PM2.5 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.

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 PM2.5 equivalent mass concentration near the surface. The performance of MERRA2 calculated PM2.5 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 PM2.5 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, PM2.5 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 PM2.5 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 PM2.5 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 PM2.5 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 PM2.5 follows diurnal, day-to-day, and seasonal cycles as observed by ground monitors.
The MLA underestimates PM2.5 under high PM2.5 concentration (>80), 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.

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 PM2.5 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 models.

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.

REFERENCES


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


Matter in a Growing Asian Economy: The Case of Bangkok Metropolitan Region.

Atmosphere 10. https://doi.org/10.3390/atmos10050227


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


Gupta et al., 2021, Thai Air


https://doi.org/10.1029/2020EA001599
Table 1. Statistical performance parameters for inter-comparison of MERRA2 and MERRA2_ML with observed PM2.5 at hourly time scale. The scatter plot for these data are shown in Figure 3.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MERRA2</td>
</tr>
<tr>
<td>N</td>
<td>266094</td>
</tr>
<tr>
<td>R</td>
<td>0.53</td>
</tr>
<tr>
<td>Bias</td>
<td>-3.8</td>
</tr>
<tr>
<td>RMSE</td>
<td>16.7</td>
</tr>
<tr>
<td>Slope (M)</td>
<td>0.38</td>
</tr>
<tr>
<td>Intercept (c)</td>
<td>9.0</td>
</tr>
</tbody>
</table>
Fig 1. Schematic of data integration and machine learning algorithm processes.
Fig. 2. Frequency distribution of input and output parameter for the machine learning algorithm.
Fig. 3. The comparison between hourly observed (x-axis) and estimated PM2.5 (y-axis). The top panel is for MERRA2 derived PM2.5 whereas bottom panel is for MLA retrieved PM2.5 using MERRA2 aerosols and meteorology as input. The statistical parameter for this analysis is presented in table 1.
Fig. 4. The regional performance of MERRA2_ML PM2.5 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).
Fig. 5. The diagnostic (top pane, bias as function of observed PM2.5) and prognostic (bottom, bias as function of MERRA2_ML PM2.5) errors for hourly MERRA2_ML PM2.5. The mean bias (y-axis) is calculated as (estimated – observed) PM2.5. The red color shows bin average values whereas green shows median values.