

Hourly and Daily PM_{2.5} Estimations using MERRA-2: A Machine Learning Approach

Alqamah Sayeed^{*1,2}, Paul Lin¹, Pawan Gupta^{1,2}, Nhu Nguyen Minh Tran³, Virginie Buchard^{4,5}, Sundar Christopher³

¹Universities Space Research Association (USRA), Huntsville, AL, USA

²NASA-Marshall Space Flight Center, Huntsville, AL, USA

³The University of Alabama in Huntsville, Huntsville, AL, USA

⁴University of Maryland Baltimore County, Baltimore, MD, USA

⁵NASA Goddard Space Flight Center, Greenbelt, MD, USA

*corresponding author, alqamah.sayeed@nasa.gov

Abstract

Health and environmental hazards related to high pollutant concentrations have become a serious issue from the perspectives of public policy and human health. The objective of this research is to improve the estimation of grid-wise PM_{2.5}, a criteria pollutant, by reducing systematic bias in estimating PM_{2.5} empirically from speciation provided by MERRA-2 using a ML approach. We present a unique application of machine learning (ML) for estimating hourly PM_{2.5} concentrations at grid points of Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2). The model was trained using various meteorological parameters and aerosol species simulated by MERRA-2 and ground measurements from Environmental Protection Agency (EPA) air quality system (AQS) stations. monitors. The ML approach significantly improved performance and reduced mean bias in the 0-10 $\mu\text{g m}^{-3}$ range. We also used the Random Forest ML model for each EPA region using one year of collocated datasets. The resulting ML models for each EPA region were validated and the aggregate data set has a Pearson correlation of 0.88 (RMSE = 4.8 $\mu\text{g m}^{-3}$) and 0.82 (RMSE = 5.8 $\mu\text{g m}^{-3}$) for training and testing, respectively. The correlation (and RMSE) increased to 0.89 (4.0), 0.95 (1.6), 0.94 (1.1) for daily, monthly, and yearly average comparisons. The results from initial implementation of the ML model for global region are encouraging but require more research and development to overcome challenges associated with data gaps in many parts of the world.

Plain Language Summary:

PM_{2.5} is one of the most important pollutants that is monitored globally since it is both a serious health and environment hazard. The monitoring helps in understanding the air quality and provide advisory to public. This advisory becomes more important in the scenarios of natural or anthropogenic events where there is a sudden increase in its concentration. Although a ground-based monitor can detect a high episode, its spatial extent, however, cannot be detected. In order to understand the spatial spread, numerical model outputs like MERRA-2 provide a good estimation of various meteorology and aerosols at horizontally gridded location. However, due to coarse resolution and averaging of large grid box, results in certain uncertainties which leads to biases. In this study we implemented a machine learning model to provide better estimation of PM_{2.5} at grid-level using MERRA-2 as base model. The developed ML model estimated the PM_{2.5} with good accuracy and very low biases.

Keywords: ML, Random Forest, grided estimation, MERRA-2, PM_{2.5}

1. Introduction

Deteriorating air quality has emerged as a major concern of the twentieth century since it impacts global climate change and health hazards, and significantly influences socioeconomic policies. Air pollution claims three times more lives than road traffic accidents globally (Myllyvirta, 2020; World Health Organization, 2018). In addition, a recent report from the Centre for Research on Energy and Clean Air (CERA) estimates economic costs resulting from air pollution to be 2.9 trillion USD (Myllyvirta, 2020). According to the report, in 2018, PM_{2.5} (mass concentrations of fine particle less than 2.5 μm in aerodynamic diameter) caused 4.5 million premature deaths and was responsible for 1.8 billion days of work absence. In another report, the EPA (US EPA, 2015) estimated that after implementation of the Clean Air Act of 1970, 2.3 million premature deaths were avoided in the next 40 years. The report also estimated that the cost to benefit ratio was \$1 to \$30 for moderate estimates; the high estimate was \$1 to \$90, and the low estimate was \$1 to \$3. These estimates imply that economic benefits exceeded costs by at least a factor of 3. Therefore, developing policies specific to environmental pollutants is of paramount importance, and estimating the quantity and impact of pollutants is a necessary step.

With recent advancements in measurement techniques, remote sensing observations, ground and surface observations, determining the atmosphere's composition (physical and chemical) and its constituents (ozone, SO₂, PM_{2.5}, NO_x etc.) has become easier. Despite these advancements, generating a spatially complete dataset is still challenging. It is also difficult to collect in-situ measurements in remote geographical locations. Deploying a uniformly and densely distributed network of stations is not cost effective, and such a network would be demanding to manage. Satellite Remote sensing has its own limitations such as measuring under cloud cover (Christopher & Gupta, 2010), retrieving data in snowy regions, scanning complex surfaces, and detecting heavy aerosols layers near sources (Hoff & Christopher, 2009). Significant effort and resources have been devoted to developing systems that improve grid-wise estimation of atmospheric constituents and the state of the atmosphere. However, these systems have been limited by the quality and amount of data (Ghahremanloo et al., 2021) and by the performance of numerical and/or chemical transport models (Pouyaei et al., 2020). Addressing these limitations incorporates both in-situ measurements and remote sensing products into data assimilation techniques (Bocquet et al., 2019; Jung et al., 2019), which are limited by the availability of the respective data products. When combined with numerical modeling and data assimilation, these measurement methods provide the true state of the atmosphere at any given point in space and time. Over several decades, studies have reported significant advances in measuring and assessing surface features (Mulla, 2013) as well as forecasting and managing air quality (Mhawish et al., 2018). Satellite remote sensing data sets have contributed essential data pertaining to the global distribution (Christopher and Gupta, 2020; Lee et al., 2016; Martin, 2008), evolution (Q. Zhang et al., 2012), and transport (Kim et al., 2017; Y. Wu et al., 2018) of atmospheric pollutants. As such, methods to estimate surface concentrations of PM_{2.5} between in-situ measurement stations are essential to address the accuracy limitations of model simulations and spatial limitations of air quality stations.

Several reanalysis techniques like Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) developed by NASA's Global Modeling and Assimilation Office (GMAO) provide grid-wise concentrations of various aerosols contributing to the concentration of PM from 1980 to current. The concentrations of PM_{2.5} and PM₁₀ can be estimated using empirical relationships from the aerosol products of the reanalysis (Buchard et al., 2016b). The reanalysis products provide the concentrations of the particulates (speciation). These

particulate concentrations can be utilized to estimate $PM_{2.5}$ (equation 1) and PM_{10} concentrations empirically. Many studies have utilized various methods to impute missing data by using local linear histogram-matching (Storey et al., 2005), regression models (Eslami, Salman, et al., 2019; Ghahremanloo et al., 2021; Rulloni et al., 2012), inpainting algorithms (Bugeau et al., 2010; Liu et al., 2018), geostatistical approaches such as Kriging (Yu et al., 2011; C. Zhang et al., 2007), and deep learning algorithms (Li et al., 2017; Q. Zhang et al., 2018). Deep learning models (Y. LeCun et al., 2015) such as convolutional neural networks (CNN) show significant promise in addressing the limitations of missing data by modeling high-level abstractions within datasets (Bengio, 2009; Deng, 2014). Furthermore, CNN (Krizhevsky et al., 2017) are among the most successful and widely used approaches (Lecun & Bengio, 1995) in forecasting (Eslami, Choi, et al., 2019; Lops et al., 2019; Sayeed, Lops, et al., 2021; C.-Y. Zhang et al., 2015), speech recognition (Mikolov et al., 2011), and imputation (Li et al., 2017; Q. Zhang et al., 2018). Gupta et al., (2021) showed that among various Machine Learning (ML) models, Random Forest estimated $PM_{2.5}$ hourly concentration with a correlation of 0.95 and mean bias (MB) of $0.03 \mu g m^{-3}$. They used various meteorology and aerosol diagnostic products from MERRA-2.

Despite the efforts made, establishing a reliable and spatially complete dataset is still a challenge, primarily due to the lack of adequate ground-based observations, missing data due to cloud cover in passive satellite remote-sensing, and uncertainties in methods and outputs from methods. In this study we aim to address these issues and develop a ML model that can estimate hourly $PM_{2.5}$ at MERRA-2 grid resolution by training data-driven models for the United States. To do this, several ML models were developed similar to the Gupta et al., (2021) model. The models were trained using in-situ measurements and MERRA-2 generated meteorology and aerosol diagnostic parameters. The developed models were then evaluated and selected based on their statistical performance. Once an ML model was selected, two different analyses were performed: regional analysis, wherein the selected model was optimized for the regions defined by the EPA, and comparative analysis with empirically estimated $PM_{2.5}$ from MERRA-2.

2. Data and Methods

The proposed methodology focuses on the estimation of the hourly $PM_{2.5}$ concentration based on the hourly time-averaged meteorology and aerosol products from MERRA-2. Figure 1 shows the process flow diagram of the proposed methodology. The methodology includes data retrieval from MERRA-2 and ground stations, spatial and temporal collocation, and model training. To determine the best model, various ML models were trained and evaluated based on several statistical parameters. The ML models were trained using meteorology and aerosol diagnostics from MERRA-2 reanalysis runs as input and spatio-temporal collocated ground measurements as output. The detail methodology is discussed in further sections.

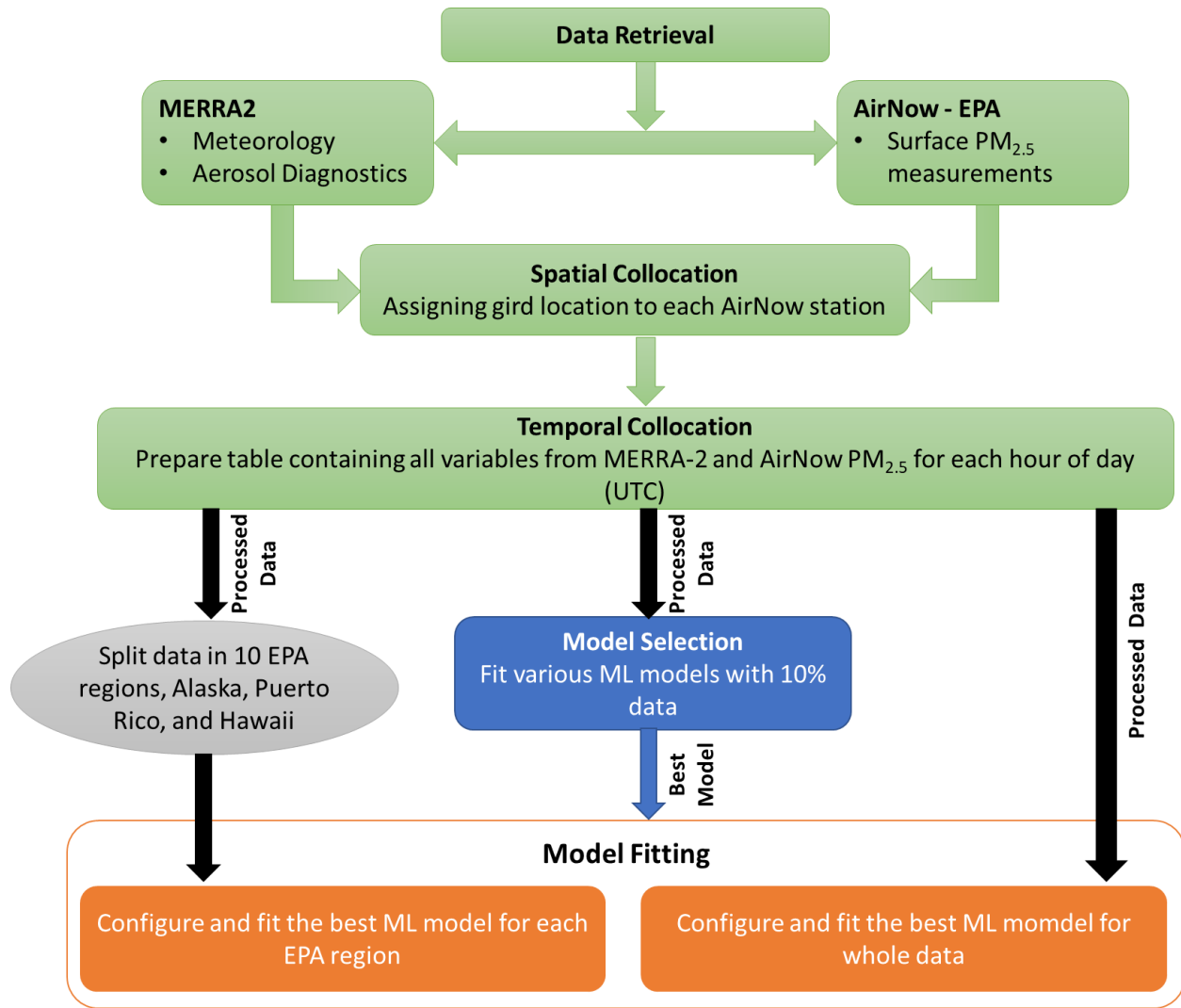


Figure 1: Schematic diagram of the methodology used for estimation of $PM_{2.5}$. The color code of boxes represents the major steps in model development. The boxes in green represents data pre-processing; the blue boxes represent model selection, and the orange box represents model fitting.

2.1. Models

In this study, several ML models and a numerical model based on reanalysis were used and are described in detail in the following sections.

2.1.1. MERRA-2

MERRA-2 is a reanalysis model for the atmosphere developed by NASA's GMAO (Gelaro et al., 2017) and provide various meteorology and aerosol products from 1980 to current. MERRA-2 uses the atmospheric model from Goddard Earth Observing System (GEOS; Molod et al., 2015; Rienecker et al., 2011) and the grid point statistical interpolation (GSI) analysis scheme (Kleist, Parrish, Derber, Treadon, Errico, et al., 2009; Kleist, Parrish, Derber, Treadon, Wu, et al., 2009; W.-S. Wu et al., 2002). The model uses cube-shaped horizontal discretization with approximate resolution of $0.5^\circ \times 0.625^\circ$ and has 72 vertical pressure layers (surface to 0.01 hPa). The MERRA-2 model also uses a 3-dimensional variational (3DVAR) data assimilation algorithm based on 6-hour update cycle of GSI for analysis. In addition, it applies first guess at appropriate time (FGAT)

for computing departures (observation - background). These analyses are further applied to correct the background state using the incremental analysis update method (Bloom et al., 1996).

MERRA-2 also produces 3-hourly analyses of gridded aerosol diagnostics on a global scale by assimilating bias corrected-AOD at 550nm. The AOD values are retrieved from various satellite sources including Moderate Resolution Imaging Spectroradiometer (MODIS), Advanced Very High Resolution Radiometer (AVHRR), and Multi-angle Imaging SpectroRadiometer (MISR) and the surface-based AErosol RObotic NETwork (AERONET) for data assimilation in MERRA-2 (Randles et al., 2017). It is important to note that due to various data availability issues, specific sensor's data assimilation into MERRA2 varies by years. For example, MODIS data assimilation were only after year 2000 and MISR data assimilation were discontinued after year 2015 (Randles et al., 2017). The aerosol components (dust, sea salt, black carbon, organic carbon, and sulfate) are derived from simulating the Goddard Chemistry Aerosol Radiation and Transport (GOCART) model (Chin et al., 2002; Colarco et al., 2010) coupled with the GEOS atmospheric model. Since MERRA-2 provides both meteorological parameters and aerosol diagnostic species at a comparatively fine resolution of approximately 50km× 65km (0.5° latitude and 0.625° longitude) on a global scale, we used these derived products for this study. In addition, PM_{2.5} can be calculated using aerosol diagnostic products of MERRA-2, shown in equation 1 below (Buchard et al., 2016b; Malm et al., 1994, 2011).

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

The equation shows the empirical relationship among the concentrations of PM_{2.5}, black carbon (BC), organic carbon (OC), sulfate (SO₄), dust (DUST_{2.5}; size < 2.5µm) and sea salt (SS_{2.5}; size < 2.5µm). In this study, two different MERRA-2 products were used: a) time-averaged hourly meteorology diagnostics at various levels (M2T1NXSLV or tavg1_2d_slv_Nx) and b) time-averaged hourly aerosol diagnostics at the surface and column AOD (M2T1NXAER or tavg1_2d_aer_Nx).

2.1.2. Machine Learning Models

From equation 1, the PM_{2.5} can be estimated at grid level, but the relationship has several missing terms, like nitrate and ammonium concentrations, (Buchard et al., 2016b, 2017) that may constitute the total PM_{2.5}. Also, the equation assumes that OC, BC, and SO₄ sizes are less than 2.5 µm. Due to these issues and model uncertainties, biases (low or high) can be introduced in PM_{2.5} estimations. The semi-volatile and volatile species can bias PM_{2.5} toward low or it may be high because of size growth due to soluble organic/inorganic species (Malm et al., 2011). Since ML models can help reduce systematic biases by handling complex non-linear relationships between meteorology and surface aerosols, we compared several regression and ML models as discussed below for estimating hourly PM_{2.5}.

The regression models used in this study were Ordinary Least Square (OLS), Ridge, and Lasso regression. The ML models used were Stochastic Gradient Descent (SGD), k-nearest neighbor (KNN), Adaptive Boosting (AdaBoost), Gradient Boost (GB), Extreme Gradient Boost (XGB), Support Vector Machine (SVM), and Random Forest (RF). The OLS model minimizes the residual sum of squares between inputs and output to fit a linear model with coefficients and intercepts. Alternatively, Ridge and Lasso Regression regularize model by imposing penalty (Friedman et al., 2010; Koh et al., 2007; Rifkin and Lippert, 2007) In SGD, a linear model is fitted using an SGD learning method with penalties to minimize the loss function (Y. A. LeCun et al., 2012).

KNN estimates a target (output) by interpolating the nearest neighbors of the target in the training set. In boosting methods like AdaBoost, GB, and XGB, the process depends on boosting the performance of weak learners, or the decision trees with single splits. In AdaBoost, more weight is given to observations that are difficult to classify. In the GB method, weak learners are boosted by optimization (minimizing loss) using gradient descent-like methods. XGB is an efficient and computationally fast implementation of the GB method. SVM is a supervised learning technique in which the model makes a discrete estimation based on best fit. The best fit line represents a hyperplane containing the maximum number of points. Unlike other regression models, SVM doesn't try to minimize the loss function. Instead, it attempts to predict the best fit based on the distance between the hyperplane and boundary line. RF is a decision tree-based model that ensembles various models (trees) for estimations. It fits several decision trees on various sub-samples of a dataset and improves accuracy by averaging and controlling overfitting.

2.2. Data and Pre-processing

The models were trained for the Continental United States (CONUS), Alaska, Puerto Rico, and Hawaii. Figure 2 shows the number of valid pairs of spatiotemporally collocated hourly observed, and MERRA-2 parameters at $PM_{2.5}$ monitoring stations (AirNow). The reanalysis data was obtained from

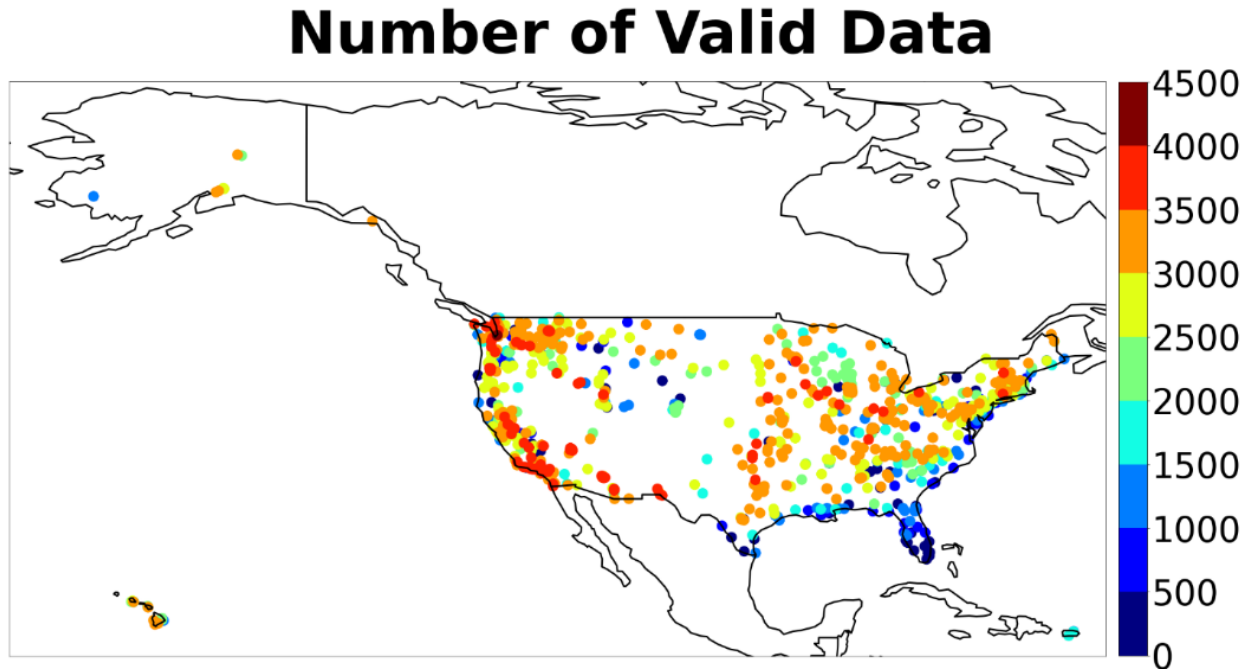


Figure 2: Locations of $PM_{2.5}$ measurement stations in the United States binned by number of valid observations in the year 2018. Each circle represents the location of a measurement station, and the color represents the number of valid observations in the year 2018.

MERRA-2 for hourly averaged meteorology and aerosol components for the year 2018. Table 1 lists all the parameters from MERRA-2 used in this study. In addition, the parameters listed in Table 1, sun-earth distance, solar zenith angle, latitude and longitude were also used as the input parameters for the model. These additional parameters accounts for change in season (sun-earth distance), diurnal cycle (solar zenith angle) and geographical location (latitude/longitude) in ML model. The parameters were selected based on a previous study by Gupta et al., (2021). The

importance of meteorology factors such as temperature, humidity, surface condition, and vegetation is well known and critical in the formation and transport of aerosols via pressure, wind speed and direction (Gupta & Christopher, 2009; Pandis, 2004; Seinfeld & Pandis, 2016). The aerosols that represent surface PM_{2.5} were selected from the MERRA-2 aerosol diagnostic product. Additionally, to represent the total aerosols in the atmospheric column, total aerosol optical thickness (AOT) measured at 550nm from MERRA-2 was selected.

Table 1: List of parameters from MERRA-2 reanalysis used as input to train ML models.

<i>Name</i>	<i>Description</i>	<i>Units</i>
Meteorology		
PS	Surface Pressure	Pa
Q500	Specific Humidity at 500 hPa	kg kg ⁻¹
Q850	Specific Humidity at 850 hPa	kg kg ⁻¹
T850	Temperature at 850 hPa	K
T500	Temperature at 500 hPa	K
T10m	Temperature at 10m	K
QV10m	Specific Humidity	kg kg ⁻¹
WIND	Windspeed at 10m	m s ⁻¹
Aerosols		
DUSMASS25	Dust Surface Mass Concentration - PM _{2.5}	µg m ⁻³
OCSMASS	Organic Carbon Surface Mass Concentration	µg m ⁻³
SO2SMASS	SO ₂ Surface Mass Concentration	µg m ⁻³
SO4SMASS	SO ₄ Surface Mass Concentration	µg m ⁻³
TOTEXTTAU	Total Aerosol Extinction AOT [550 nm]	unitless
BCSMASS	Black Carbon Surface Mass Concentration	µg m ⁻³
SSSMASS25	Sea Salt Surface Mass Concentration - PM _{2.5}	µg m ⁻³

Hourly concentration surface measurements of PM_{2.5} for 2018 were obtained from the AirNow network managed by the EPA. Several state, city, and county level data collection agencies share their data of various air quality parameters to the EPA which prepares, and quality checks the data. For this study, 882 unique monitoring sites were selected based on collocated data availability. The spatial collocation is performed by selecting the nearest MERRA-2 grid to the EPA monitoring station by spherical distance (< 38.5 kms; Gupta et al., 2021). Due to the coarse resolution of MERRA-2, some of the stations were assigned the same grid point. Temporal collocation was accomplished by selecting the same hour of data from both EPA monitor and MERRA-2 outputs. The final integrated datasets included 2,128,126 valid collocated data points. After collocating data, multiple ML models were trained with only 10% of data points which were randomly selected (212,813 of 2,128,126 valid data points). The data was further split into a 70:30 ratio for model fitting and validation. Several regression and ML models were trained with 70%

of the selected data (148969), and the remaining 30% (63844) was used for testing. The reduced sample size was only used to select the best-performing algorithm. The most successful ML model was evaluated further using full datasets.

3. Results and Discussion

To evaluate the models discussed in this study, a 10-fold cross validation algorithm was used. In this algorithm, 10 different sets of 9:1 train-test splits were created such that each set was unique, and when all the “test” samples in the 10 sets were combined gives the whole dataset. Each set was evaluated for various statistical performance parameters like root mean squared error (RMSE), mean bias (MB), Pearson linear correlation coefficient (R), slope of the fitted line, percentage error, and computational time. In addition to comparing the ML models, the performance was also evaluated compared to MERRA-2 estimated PM_{2.5} (Eq.1).

3.1. Evaluation of ML Models

All the ML models were trained and evaluated on several statistical parameters. Table 2 lists ML performance statistics for correlation, MB, RMSE, and computational time. The regression models (OLS, Ridge, Lasso) were fitted within a fraction of second and had similar RMSE and MB compared to other methods, but these models demonstrated very low correlation and high percentage error. KNN and SVM had low percentage error and better correlation than regression methods, but the computational time was much longer. GB and XGB had high percentage error and correlations of 0.66 for training and 0.61 for testing data. Adaboost was the least successful ML model. RF exhibited the highest correlations: 0.97 for training and 0.69 for testing. RF also

Table 2: Performance metrics of all ML models used in the study.

	Slope of fit		Correlation (r)		Mean Bias (MB) $\mu\text{g m}^{-3}$		RMSE $\mu\text{g m}^{-3}$		Time (sec)
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	
OLS	0.22	0.22	0.47	0.47	0.00	-0.03	8.78	8.75	0.19
Ridge	0.22	0.22	0.47	0.47	0.00	-0.03	8.78	8.75	0.09
Lasso	0.12	0.13	0.41	0.41	0.00	-0.01	9.13	9.07	0.79
SGD	0.26	0.26	0.46	0.46	0.05	0.02	8.88	8.86	1.34
KNN	0.59	0.46	0.78	0.63	0.01	-0.03	6.18	7.78	1058.99
SVM	0.26	0.24	0.60	0.57	-1.29	-1.33	8.24	8.34	3887.67
Adaboost	0.60	0.55	0.31	0.29	41.30	41.25	45.32	45.35	41.67
GB	0.38	0.35	0.66	0.61	0.00	-0.04	7.50	7.83	124.94
XGB	0.37	0.34	0.66	0.61	0.00	-0.05	7.55	7.82	22.85
RF	0.82	0.49	0.97	0.69	0.09	0.20	2.61	7.15	517.70

returned low RMSE of 2.61 $\mu\text{g m}^{-3}$ and 7.15 $\mu\text{g m}^{-3}$ for training and testing, respectively. The computation time was less than 10 minutes, which was higher than most models, but RF outperformed all other ML models in other statistical metrics. The linear models like; OLS, Ridge, Lasso and SGD; it was expected to have a lower performance metrics due to non-linearity of PM_{2.5}.

RF performed better than boosting methods like Adaboost, gradient boost and XG, because it tries to reduce the variance instead of bias, which in the case of PM_{2.5} (having diurnal variation) improved the overall accuracy. Since, RF model is an ensemble of multiple predictive decision tree, it effectively reduces biases in estimation and hence have improved the performance over other techniques. Although, there are some inconsistencies in performance between training and testing in RF models, but 10-fold cross validation further improved the consistency (table 3). Therefore, for this study, we selected RF for further evaluations.

Once the RF model was selected, 10-fold validation was performed with all valid data (2,128,126 data points). From this data, 10 sets (folds) of training and testing with a 9:1 split were prepared. Each testing set contained a unique amount of data and, when combined, included all data. Table 3 lists the performance statistics of all 10-folds. All folds performed similarly on all evaluation metrics, suggesting that the model was robust and could be used for any dataset combinations.

Table 3: Performance of all 10-folds on testing sets.

Fold	Number of Testing Data	Slope of fit	Correlation (r)	Mean Bias ($\mu\text{g m}^{-3}$)	RMSE ($\mu\text{g m}^{-3}$)
1	212818	0.65	0.83	0.01183	5.53
2	212817	0.59	0.80	-0.01882	6.28
3	212815	0.63	0.82	-0.00564	5.73
4	212813	0.63	0.83	0.00406	5.70
5	212813	0.63	0.83	0.02189	5.61
6	212813	0.64	0.83	-0.00492	5.56
7	212812	0.64	0.83	-0.01886	5.51
8	212811	0.63	0.82	-0.01508	5.67
9	212808	0.64	0.82	0.02798	5.54
10	212806	0.62	0.81	-0.00775	5.83
Average	2128126 (Total)	0.63	0.82	-0.00053	5.70

3.2. Model Performance on EPA regions

In the next phase of RF model evaluation, the whole dataset was divided into 10 geographical regions assigned by the EPA (US EPA, 2020). Supplementary figure S1 shows each EPA region. Although Puerto Rico, Hawaii, and Alaska were assigned to regions 2, 9, and 10 respectively, they were evaluated separately as they are not part of CONUS. The RF model was trained with 70% of the data points from each region separately and tested on the remaining 30% of data with six different combinations of number of estimators (N), maximum depth (D), and minimum sample leaf (m), which are sklearn RF parameters. The best combination for each region was selected to further evaluate the RF model. Table 4 and figure 3 show the spatial distribution of correlation, mean bias, and RMSE for RF model performance against those measured at ground locations on test dataset. The correlations are higher for stations in western US but with high value of RMSE suggesting higher PM_{2.5} concentration compared to eastern US. Regions 10, 9, Hawaii, and Alaska in the west performed the best in terms of correlations, earning greater than 0.90 for training and greater than 0.82 for testing. These regions have high number of valid data points overall and this helps in better training of model due to uniform distribution of data-points both spatially and temporally. Additionally, even though Hawaii and Alaska have lesser number of data-points it was amongst the best performing region due to uniform PM_{2.5} variability throughout the region. Region 9, which includes California, Nevada, and Arizona, had the highest RMSE due to greater PM_{2.5}

concentrations ($10.88 \pm 14.06 \mu\text{g m}^{-3}$) compared to other regions ($8.45 \pm 8.62 \mu\text{g m}^{-3}$). Supplementary figure S2 shows the frequency distribution of observed $\text{PM}_{2.5}$ by region. Puerto Rico and region 8, which includes Colorado, Utah, Montana, Wyoming, and North and South Dakota, have correlations ~ 0.88 for training and ~ 0.80 for testing. In the Northeast, regions 1 and 2 had training correlations of 0.76 and 0.77, respectively, and about ~ 0.70 for testing. The lowest-performing regions were from the Midwest (region 7: Kansas City) and Southeast (region 4: Atlanta). The performance of the ML model in estimating primary and secondary pollutants (like ozone, nitrogen dioxide, PM, etc.) is affected by several meteorological and anthropogenic factors such as urbanization, diurnal variability, wind speed and direction, and boundary layer height (vertical mixing) (Sayeed, Choi, et al., 2021). Since each EPA region in the US has its distinct anthropogenic (emission, urbanization etc.), geographic (geo location, elevation etc.) and meteorological (seasonality, diurnal variability) features and characteristics, it is challenging for a generalized model to have a similar performance in all regions.

Table 4: Regional comparison of Random Forest model with best set of training parameters for each region.

EPA Regions	Number of valid data in region		Correlation		Mean Bias ($\mu\text{g m}^{-3}$)		RMSE ($\mu\text{g m}^{-3}$)		Number of Stations
	Train	Test	Train	Test	Train	Test	Train	Test	
Region 1: Boston	81544	34948	0.76	0.70	-0.01	0.01	3.25	3.51	50
Region 2: New York City	45187	19366	0.77	0.69	0.00	-0.01	3.93	4.36	37
Region 3: Philadelphia	95226	40812	0.70	0.63	0.00	0.00	4.29	4.56	65
Region 4: Atlanta	174515	74793	0.65	0.59	0.00	0.00	4.20	4.34	129
Region 5: Chicago	214164	91785	0.67	0.61	0.00	0.00	4.68	5.00	115
Region 6: Dallas	105678	45291	0.71	0.66	0.00	0.01	4.48	4.62	63
Region 7: Kansas City	75644	32419	0.69	0.61	0.00	0.01	4.62	4.91	36
Region 8: Denver	85176	36505	0.87	0.78	0.00	0.04	4.71	5.80	56
Region 9: San Francisco	296181	126936	0.91	0.85	0.00	0.03	5.92	7.52	151
Region 10: Seattle	276414	118464	0.94	0.87	0.00	-0.01	4.84	6.78	154
Alaska	14932	6400	0.95	0.82	0.03	0.14	3.05	4.88	8
Hawaii	22054	9452	0.94	0.82	0.01	0.01	2.74	4.67	13
Puerto Rico	2968	1272	0.89	0.81	0.09	0.29	4.92	6.67	3

It is also evident from figure 2 and supplementary figures S3 and S4 that regions with the larger concentrations range of $\text{PM}_{2.5}$ performed better than regions with lower concentrations ranges. This suggests that the ML model is sensitive to concentration variability, performing better in regions

with high $\text{PM}_{2.5}$ compared to regions with uniform $\text{PM}_{2.5}$. This is possibly due to the skewness of data (number of data-points with low concentrations far exceeds the high concentrations), while the higher representations of highs make the performance better, its lower representation makes the performance comparatively poor. This is an expected behavior of regression or ML based models where data distribution often force model to estimate mean better than extreme values (low and high) (Gupta et al., 2021; Gupta & Christopher, 2009; Ma et al., 2021). It was also noted that, in regions with less variability, the RF model performed better in areas with less than 30,000 data points, suggesting that adding more data to a less variable region makes the model susceptible to errors possibly due to overfitting.

Next, we present results of inter-comparison of training and testing separately while combining data from all the regions (i.e., Table 4). Figure 4a shows density scatter plot of aggregated data points from all the regions together. The left panel shows training datasets while right panel is for testing datasets. The scatter plot demonstrates overall good correlations with consistent performance between training ($r=0.88$) and testing ($r=0.82$). The mean bias is close to zero in both datasets with RMSE changes from $4.8 \mu\text{g m}^{-3}$ in training to $5.8 \mu\text{g m}^{-3}$ in testing datasets. The slope values of (~ 0.7 and 0.62) implies overall underestimation of $\text{PM}_{2.5}$ by ML models. We further evaluated the ML model

performance by combining data from training and testing and presenting as diagnosis and prognosis errors in Figure 4 (b) and (c) respectively. The complete collocated dataset was sorted according to $\text{PM}_{2.5}$, binned into bins with equal number of collocations (1000 data points), and then mean, median and standard deviation of each bin was calculated. Orange dots and line show the mean. The green dots and line show the median. The grey shaded area indicates one standard deviation of each bin. The horizontal black dotted line denotes zero difference. Positive values of bias indicate that estimated $\text{PM}_{2.5}$ is higher than observed AirNow $\text{PM}_{2.5}$ and vice-versa. The figure 4(a) shows that ML model overestimates $\text{PM}_{2.5}$ values when observed values are lower than $10 \mu\text{g m}^{-3}$ otherwise it underestimates for the remaining observed range of $\text{PM}_{2.5}$. The negative bias is more stable for values lower than $100 \mu\text{g m}^{-3}$ but then decreases exponentially for the higher $\text{PM}_{2.5}$ values. This behaviors of bias with observed $\text{PM}_{2.5}$ suggest a manifestation of data distribution with high density of low concentration values, produces better training for the ML model and

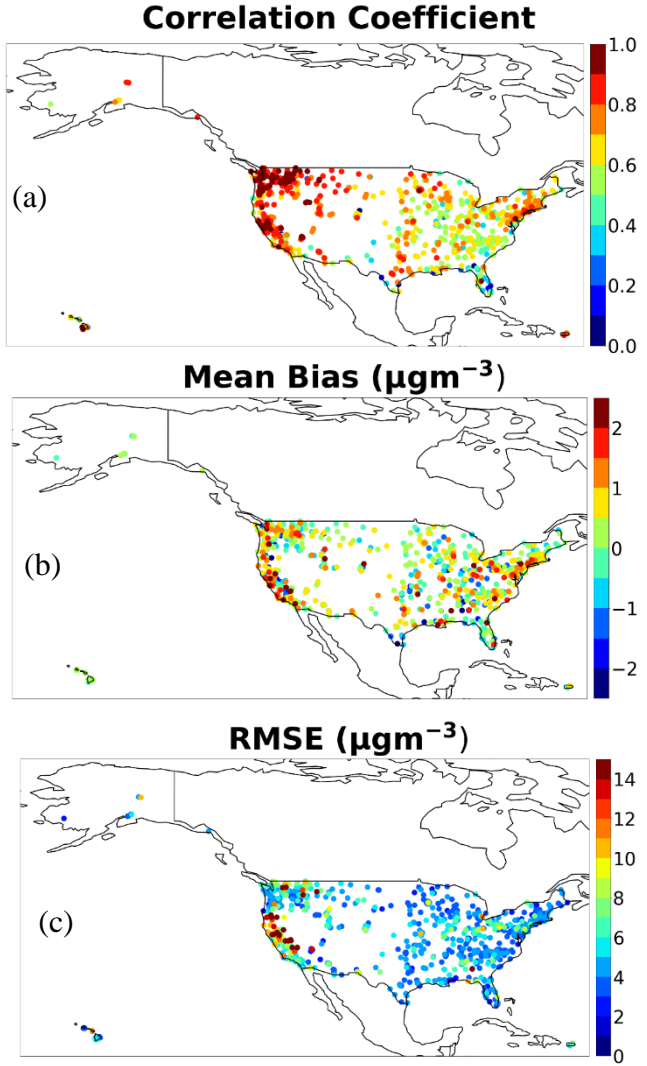


Figure 3: Regional performance of Random Forest model on test dataset; Station-wise. a) Pearson correlation of RF model trained by region; b) MB of RF model trained by region; c) RMSE of RF model trained by region.

consistent with our previous findings (Gupta et al., 2021). Figure 4 (c) show bias as a function of estimated $PM_{2.5}$ with more flat behavior for estimated $PM_{2.5}$ values smaller than $100 \mu g m^{-3}$ but then underestimates. The diagnosis and prognosis error analysis are as expected and suggests that applying more advanced data sampling methods such as oversampling (Vu et al., 2022) may require to achieve better results for complete range of observed $PM_{2.5}$.

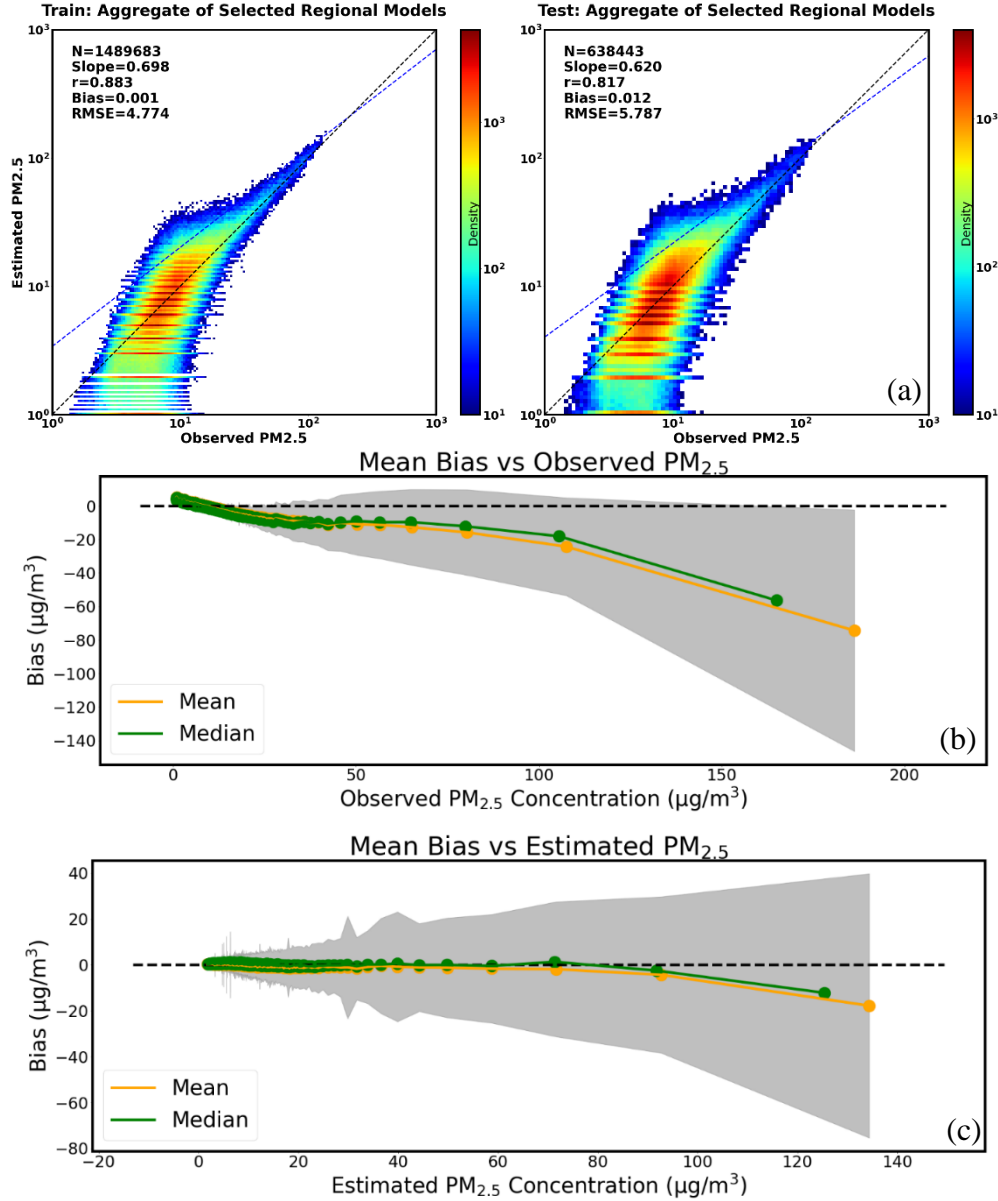


Figure 4: a) Scatter plot of estimated $PM_{2.5}$ by RF (aggregate of region models): left panel- training, right panel- testing. Black dotted line represents 1:1 line whereas, blue dotted line represents linear fit to the data; b) Mean bias ($\mu g m^{-3}$) in estimated $PM_{2.5}$ by RF (aggregate of region models) as function of observed $PM_{2.5}$. c) Mean bias ($\mu g m^{-3}$) in estimated $PM_{2.5}$ by RF (aggregate of region models) as function of estimated $PM_{2.5}$. [Bias and RMSE in $\mu g m^{-3}$]

3.3. Comparison of RF model with MERRA-2 estimates

The RF model estimates were compared to MERRA-2 estimated $PM_{2.5}$. $PM_{2.5}$ is not a direct product of MERRA-2, rather it was calculated empirically using equation 1. Additionally,

MERRA-2 has the advantage of providing data everywhere, even in places and times when observations are not available. However, MERRA-2 data is representative of a $\sim 50 \times 50 \text{ km}^2$ grid box area and has limitations due to uncertainties in emissions (Buchard et al., 2016a). Figure 5a&b shows the density scatter plot comparison of $\text{PM}_{2.5}$ estimated by MERRA-2 (equation 1) and the RF model (MERRA-2-ML) respectively. The scatter plots are plotted with aggregate of all data (training and testing combined) with observations as x-axis and estimations as y-axis. The dotted black line represents the 1:1 line. These clearly show the larger scatter in MERRA-2 $\text{PM}_{2.5}$ when comparing with ground monitors, with slope value of 1.15, mean bias of $\sim 4.1 \mu\text{g m}^{-3}$ and RMSE of $28.7 \mu\text{g m}^{-3}$. The plot shows that the MERRA-2 estimates have high positive bias with low correlation ($r=0.38$). This is also evident in supplementary figure S5; during summer and spring, MERRA-2 estimates exhibit large biases, and MERRA-2-ML successfully reduced those biases. MERRA-2 shows a better agreement with Interagency Monitoring of Protected Visual Environments (IMPROVE) sites as compared to EPA- AQS sites as IMPROVE sites are generally

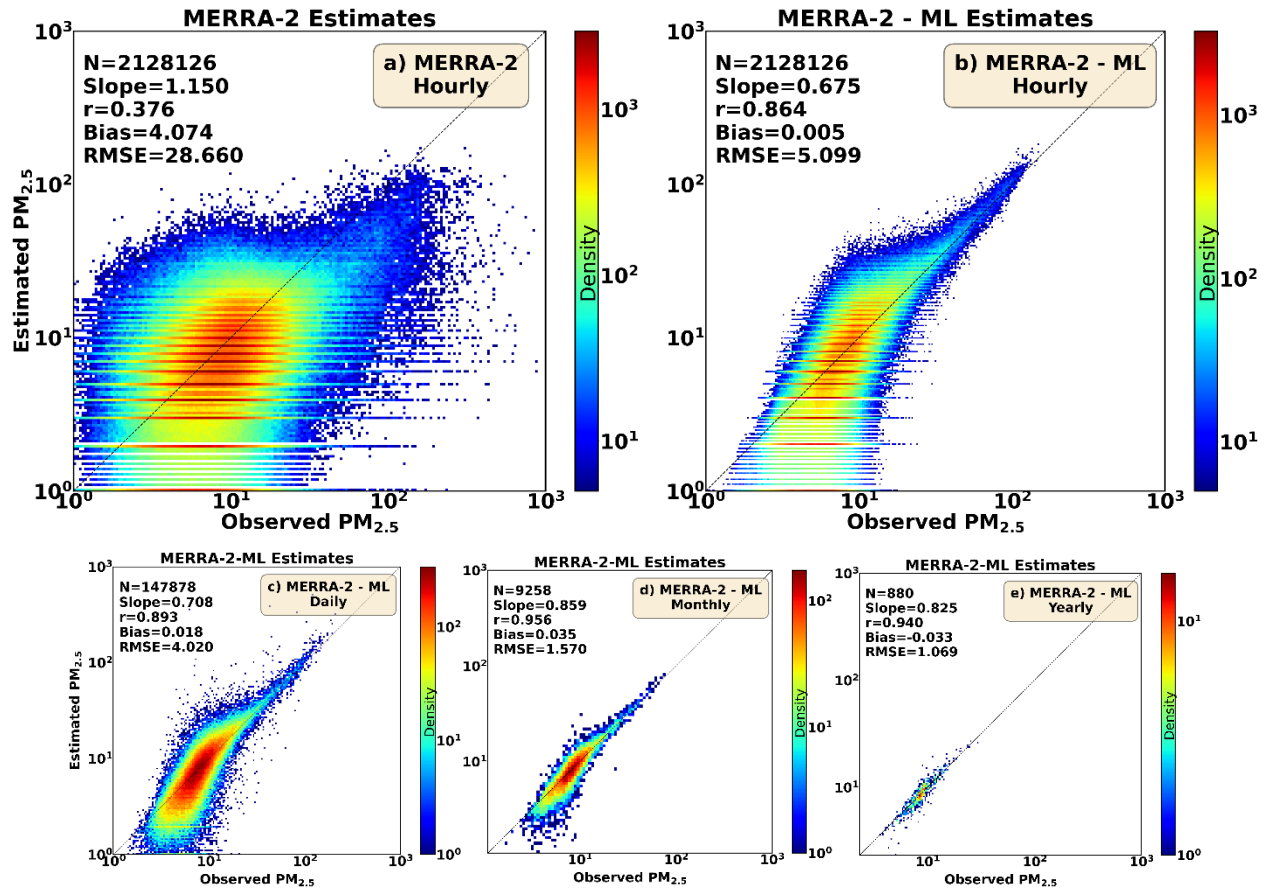


Figure 5: Scatter plot comparison of: a) MERRA-2 estimated $\text{PM}_{2.5}$; b) ML model (RF) estimated $\text{PM}_{2.5}$; c) Daily estimated $\text{PM}_{2.5}$ for MERRA-2-ML of all stations combined; d) Monthly estimated $\text{PM}_{2.5}$ for MERRA-2-ML of all stations combined and; e) Yearly estimated $\text{PM}_{2.5}$ for MERRA-2-ML of all stations combined; The color bar represents the number density. The black and blue dotted lines represent the 1:1 line and line of linear fit for the data respectively. Units are in $\mu\text{g m}^{-3}$.

located in rural remote areas having less emission variability (Buchard et al., 2016a). During fall and winter, the MERRA-2 overestimation was reduced, but the RF model still performed better. Figure 5 c, d, & e shows the scatter density plot for the RF model (MERRA-2-ML) where y-axis represents the estimation and x-axis represents the observations. The narrow spread and

concentration of high density of points near 1:1 line suggest that the estimated $PM_{2.5}$ is in good agreement with observed $PM_{2.5}$. Figure S6 shows the regional daily mean $PM_{2.5}$ concentrations that were observed and the concentrations modeled by MERRA-2 and MERRA-2-ML.

During the 2018 California wildfire season, very high $PM_{2.5}$ concentrations were observed in August through September and again in November in regions 8, 9, and 10. MERRA-2 estimated high concentrations for these periods, but the estimates exhibited large overestimations with days exceeding observations by $75 \mu g m^{-3}$ (see figure S6). In contrast, the MERRA-2-ML (RF) model did not overestimate by $6 \mu g m^{-3}$ except for four days in Puerto Rico where bias was between 6 - $13 \mu g m^{-3}$. Figure 5b shows the daily, monthly, and yearly scatter plots for observed and estimated $PM_{2.5}$. As the averaging window increases, the correlation between observations and estimations also increases. The negative bias in lower range ($<10 \mu g m^{-3}$) of observed $PM_{2.5}$ remains an issue even for larger time averages.

3.4. Assessment of Western US fires using estimated $PM_{2.5}$

To evaluate the RF model performance on the out-of-box scenarios, the model was tested by estimating hourly $PM_{2.5}$ from August 20-27, 2020. The previously trained models (using 2018 data) for each EPA region were used to estimate the hourly $PM_{2.5}$ at each grid corresponding to MERRA-2 for the CONUS. This time period was selected because, during this period, several fires were reported in the western US. Figure 6 shows a series of corrected reflectance RGB images from Suomi NPP Visible Infrared Imaging Radiometer Suite (VIIRS) satellite. The superimposed red dots are fire locations detected by VIIRS (<https://worldview.earthdata.nasa.gov/>). The second column shows observed $PM_{2.5}$ from the EPA AirNow network, and the last column includes maps of estimated $PM_{2.5}$ using MERRA-2-ML. The data corresponds to 13:00 local time for a) Aug 20, 2020, b) Aug 22, 2020, c) Aug 24, 2020, and d) Aug 26, 2020. The RGB images (column 1) shows fire locations (red dots) and smoke (grayish haze type looking feature in the image) transport across the western US. The AirNow network specifically in CA observed high $PM_{2.5}$ ($>60 \mu g m^{-3}$) values and corresponds to visible thick smoke plume seen in RGBs. It is important to note here that visible smoke in RGB can be located at any height in the atmosphere and not necessary affect surface air quality. By comparing the RGB and $PM_{2.5}$ maps, it appears that spatial patterns in estimated MERRA2-ML $PM_{2.5}$ matches with larger smoke plumes. The RF model was able to capture the movement of the plume from west to east during August 20-26, 2020. RF model results show a high concentration of $PM_{2.5}$ east and northeast of the recorded fire on Aug 20, 2020, which is also evident from the VIIRS image as haze. The plume can be seen migrating from California in the west to the northern US, crossing over the Rocky Mountains, and then moving down toward the central US. A similar rise in $PM_{2.5}$ is also seen in these regions, as estimated by the RF model. The ability of MERRA2-ML to map the smoke plumes and provide corresponding surface $PM_{2.5}$ in areas with sparse $PM_{2.5}$ ground measurements can be very useful in time of increasing wildfires in the US. The refined MERRA2-ML data can also be used to inter-compare with other regional model outputs for both analysis and forecasting.

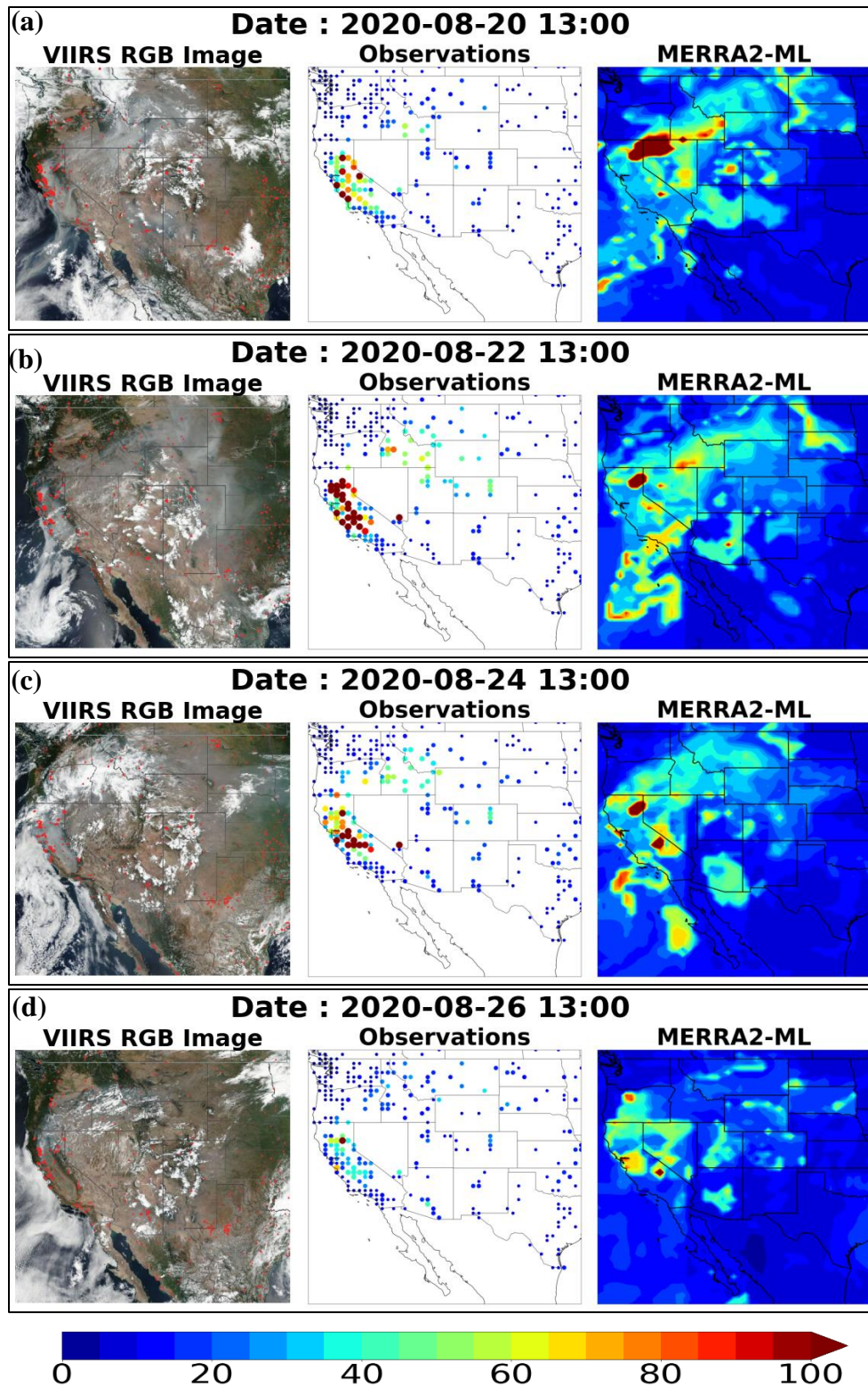


Figure 6: Comparison of VIIRS true color images with observed and estimated $PM_{2.5}$ for various days in August 2020 at 13:00 Local Time: a) Aug 20, 2020; b) Aug 22, 2020; c) Aug 24, 2020; d) Aug 26, 2020. The true color images were created using NASA Worldview (<https://worldview.earthdata.nasa.gov/>). Units for both Observations and MERRA-2 $PM_{2.5}$ are in $\mu g m^{-3}$.

4. Global Implications and Challenges

The new ML models presented here are designed to work only over the Continental United States (and perhaps other regions with similar environmental conditions). Implementing the similar ML for the global regions may be challenging and requires more extensive research development. In the CONUS we had a relatively good dense network of ground stations to work with: 800+ AirNow stations with distribution across various EPA regions. We were able to train the ML models by using relationships in input aerosols components and meteorological parameters from MERRA-2 to estimate surface $PM_{2.5}$. The lack of ground monitors, varying amount of data availability, change in aerosol type, varying meteorological conditions, and aerosol transport in other parts of the world will make accurate estimation of $PM_{2.5}$ from MERRA-2 a more complex problem. However, because the results

over CONUS have been so encouraging, we decided to train a similar RF model for entire globe using same one year of collocated datasets using more than 3000+ ground station. The ground data for global regions were collected from OpenAq platform (openaq.org). We had total six million plus collocated data points to train and test a ML model. Figure 7 presents the comparison between observed $PM_{2.5}$ with ML estimated $PM_{2.5}$ from MERRA-2 for the entire global data set, combining data for both training and testing of ML model (Random Forest). The results are surprisingly good. The global ML model has mean bias close to zero, with slope of 1.02 and RMSE of $6.4 \mu g m^{-3}$ globally. These initial results and analysis are very encouraging, but more in-depth analysis for specific regions will be required to better understand the performance and limitation of ML approach before applying it operationally at global scale.

5. Conclusions

In this study, we developed ML models to estimate hourly $PM_{2.5}$ using aerosol and meteorological parameters from MERRA-2 reanalysis for United States. Ten different ML algorithms were tested to select the best algorithm to accurately model the surface $PM_{2.5}$. Comparing these ML algorithms showed that the Random Forest algorithm performed best and computationally faster than other ML algorithms like SGD, KNN, Adaboost, Gradient Boost and XGB, and linear regression models like Lasso and Ridge. The RF model correlation values were ~25% better for training and ~10% better for testing compared to KNN, which was the next best model.

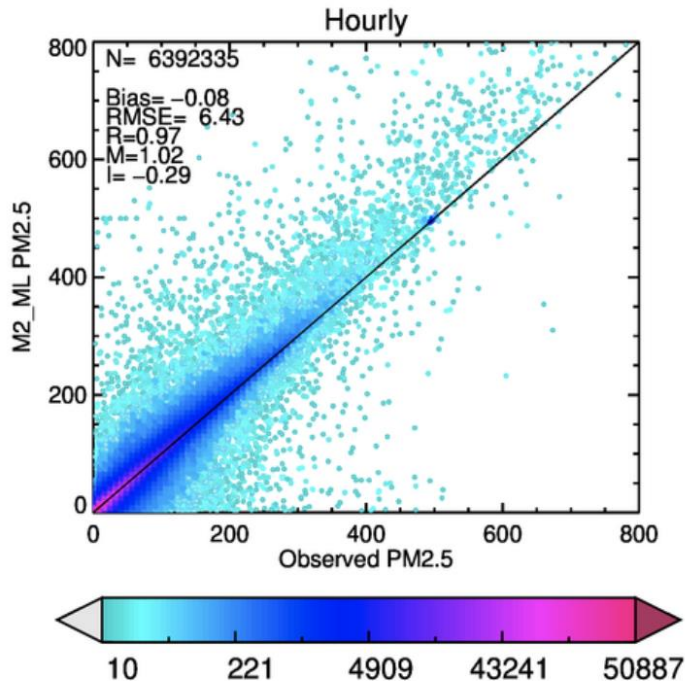


Figure 7. Density scatter plot of observed and ML estimated MERRA2 surface $PM_{2.5}$ over global region. In this plot, data from both training and testing are combined. Solid black line is 1:1 line.

The RF model was further evaluated with 10-fold cross validations. The cross validations were performed by dividing all valid data into ten sets. The sets had similar performance metrics and average correlations. Averaging all ten data sets, the overall correlation was 0.82, and the RMSE was $5.70 \mu\text{g m}^{-3}$ with mean bias close to zero. To further improve performance, separate regional RF models were developed with a combination of model parameters.

Regional RF model performance improved from east to west. The best performing regions were regions 9 and 10, Alaska, Puerto Rico, and Hawaii. While variability in observed $\text{PM}_{2.5}$ is the likely reason for favorable results in regions 9 and 10, better performance in Alaska, Puerto Rico, and Hawaii is likely due to fewer data points. Variability in large data sets produces a greater number of decision trees in the RF model which improves the performance. Attaining higher estimation accuracy for data sets with less variability can be achieved by using less data. However, using fewer data points might also induce more error in out-of-box scenarios. Therefore, larger data sets are preferable to smaller data sets despite having lower accuracy. This is one of the limitations of RF models.

The RF model was also compared with MERRA-2 empirical estimation of $\text{PM}_{2.5}$. The MERRA-2 grid estimations (larger spatial area average) largely overestimated as compared the observed $\text{PM}_{2.5}$ at point location by ground monitors. While the ML model has small and uniform bias throughout the day and in all seasons, MERRA-2 exhibits comparatively higher overestimation between 07:00 UTC and 13:00 UTC. MERRA-2 also produces a slightly lower mean of overestimation during winters, but all other seasons the mean was greater than 0. Since the ML model was better at estimating $\text{PM}_{2.5}$ than empirical calculations from MERRA-2, it can be extended to other years and grids of MERRA-2 reanalysis while taking account in change in assimilation datasets into MERRA2 datasets. This could build a spatiotemporal trend for further research and model development. Another application of such algorithm could be bias-correction of $\text{PM}_{2.5}$ forecasts from the GEOS-FP model at regional scales (Gupta et al., 2021).

To further determine the robustness of the developed model, a special case of western US fires during August 20-26, 2020, was studied. The regionally trained model was used to estimate $\text{PM}_{2.5}$ at each MERRA-2 grid location for the period. The estimated values accurately represented the plume seen in VIIRS true color images and observed $\text{PM}_{2.5}$, suggesting the RF model can perform well in the out-of-box scenarios as well.

Acknowledgement:

The authors would like to thank Dr. Argyro Kavvada, Program Lead, Sustainable Development Goals, Applied Sciences Program, NASA Earth Science Division and Mr. Tony Kim from NASA Marshall Space Flight Center for enabling funding through NASA-Google Partnership initiative of NASA's applied science program. The ground based $\text{PM}_{2.5}$ data were obtained from US EPA's AirNow network (<https://www.epa.gov/outdoor-air-quality-data>) to support this study. The MERRA2 data were obtained from NASA's Earth Science Data Systems (<https://earthdata.nasa.gov/>). Alqamah Sayeed, Paul Lin, Pawan Gupta, Nhu Nguyen Minh Tran, and Sundar Christopher were partially supported by the NASA's Research Opportunities in Space and Earth Science (ROSES-2020), Program Element A.38: Health and Air Quality Applied Sciences Team.

Data Availability:

The ground data used in the study are available through US EPA (<https://www.epa.gov/outdoor-air-quality-data>), the MERRA2 data available through NASA EarthData (<https://earthdata.nasa.gov/>). We also plan to provide ML estimated PM_{2.5} through NASA's Health and Air Quality Applied Science Website (<https://haqast.org/>). All the data and codes will be made available through public platform.

References:

- Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1–127. <https://doi.org/10.1561/22000000006>
- Bloom, S. C., Takacs, L. L., da Silva, A. M., & Ledvina, D. (1996). Data Assimilation Using Incremental Analysis Updates. *Monthly Weather Review*, 124(6), 1256–1271. [https://doi.org/10.1175/1520-0493\(1996\)124<1256:DAUIAU>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2)
- Bocquet, M., Brajard, J., Carrassi, A., & Bertino, L. (2019). Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*, 26(3), 143–162. <https://doi.org/10.5194/npg-26-143-2019>
- Buchard, V., da Silva, A. M., Randles, C. A., Colarco, P., Ferrare, R., Hair, J., Hostetler, C., Tackett, J., & Winker, D. (2016a). Evaluation of the surface PM_{2.5} in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. *Atmospheric Environment*, 125, 100–111. <https://doi.org/10.1016/j.atmosenv.2015.11.004>
- Buchard, V., da Silva, A. M., Randles, C. A., Colarco, P., Ferrare, R., Hair, J., Hostetler, C., Tackett, J., & Winker, D. (2016b). Evaluation of the surface PM_{2.5} in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. *Atmospheric Environment*, 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. *Journal of Climate*, 30(17), 6851–6872. <https://doi.org/10.1175/JCLI-D-16-0613.1>
- Bugeau, A., Bertalmío, M., Caselles, V., & Sapiro, G. (2010). A Comprehensive Framework for Image Inpainting. *IEEE Transactions on Image Processing*, 19(10), 2634–2645. <https://doi.org/10.1109/TIP.2010.2049240>
- 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. *Journal of the Atmospheric Sciences*, 59(3), 461–483. [https://doi.org/10.1175/1520-0469\(2002\)059<0461:TAOTFT>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<0461:TAOTFT>2.0.CO;2)
- Christopher, S. A., & Gupta, P. (2010). Satellite Remote Sensing of Particulate Matter Air Quality: The Cloud-Cover Problem. *Journal of the Air & Waste Management Association*, 60(5), 596–602. <https://doi.org/10.3155/1047-3289.60.5.596>

- Colarco, P., da Silva, A., Chin, M., & Diehl, T. (2010). Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth. *Journal of Geophysical Research*, 115(D14), D14207. <https://doi.org/10.1029/2009JD012820>
- Deng, L. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3–4), 197–387. <https://doi.org/10.1561/20000000039>
- Eslami, E., Choi, Y., Lops, Y., & Sayeed, A. (2019). A real-time hourly ozone prediction system using deep convolutional neural network. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-019-04282-x>
- Eslami, E., Salman, A. K., Choi, Y., Sayeed, A., & Lops, Y. (2019). A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-019-04287-6>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.
- Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Silva, A. M. da, Gu, W., ... Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419–5454. <https://doi.org/10.1175/JCLI-D-16-0758.1>
- Ghahremanloo, M., Choi, Y., Sayeed, A., Salman, A. K., Pan, S., & Amani, M. (2021). Estimating daily high-resolution PM_{2.5} concentrations over Texas: Machine Learning approach. *Atmospheric Environment*, 247, 118209. <https://doi.org/10.1016/j.atmosenv.2021.118209>
- Gupta, P., & Christopher, S. A. (2009). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *Journal of Geophysical Research: Atmospheres*, 114(D14). <https://doi.org/10.1029/2008JD011496>
- Gupta, P., Christopher, S. A., Wang, J., Gehrig, R., Lee, Y., & Kumar, N. (2006). Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmospheric Environment*, 40(30), 5880–5892. <https://doi.org/10.1016/j.atmosenv.2006.03.016>
- Gupta, P., Zhan, S., Mishra, V., Aekakkarakunroj, A., Markert, A., Paibong, S., & Chishtie, F. (2021). Machine Learning Algorithm for Estimating Surface PM_{2.5} in Thailand. *Aerosol and Air Quality Research*, 21(11), 210105. <https://doi.org/10.4209/aaqr.210105>
- Hoff, R. M., & Christopher, S. A. (2009). Remote Sensing of Particulate Pollution from Space: Have We Reached the Promised Land? *Journal of the Air & Waste Management Association*, 59(6), 645–675. <https://doi.org/10.3155/1047-3289.59.6.645>

- Jung, J., Souri, A. H., Wong, D. C., Lee, S., Jeon, W., Kim, J., & Choi, Y. (2019). The Impact of the Direct Effect of Aerosols on Meteorology and Air Quality Using Aerosol Optical Depth Assimilation During the KORUS-AQ Campaign. *Journal of Geophysical Research: Atmospheres*, 124(14), 8303–8319. <https://doi.org/10.1029/2019JD030641>
- Kim, H. C., Kim, S., Kim, B.-U., Jin, C.-S., Hong, S., Park, R., Son, S.-W., Bae, C., Bae, M., Song, C.-K., & Stein, A. (2017). Recent increase of surface particulate matter concentrations in the Seoul Metropolitan Area, Korea. *Scientific Reports*, 7(1). <https://doi.org/10.1038/s41598-017-05092-8>
- Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Errico, R. M., & Yang, R. (2009). Improving Incremental Balance in the GSI 3DVAR Analysis System. *Monthly Weather Review*, 137(3), 1046–1060. <https://doi.org/10.1175/2008MWR2623.1>
- Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W.-S., & Lord, S. (2009). Introduction of the GSI into the NCEP Global Data Assimilation System. *Weather and Forecasting*, 24(6), 1691–1705. <https://doi.org/10.1175/2009WAF2222201.1>
- Koh, K., Kim, S.-J., & Boyd, S. (n.d.). *An Interior-Point Method for Large-Scale l_1 -Regularized Logistic Regression*. 37.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient BackProp. *Neural Networks: Tricks of the Trade*, 9–48. https://doi.org/10.1007/978-3-642-35289-8_3
- Lecun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. *The Handbook of Brain Theory and Neural Networks*. <https://nyuscholars.nyu.edu/en/publications/convolutional-networks-for-images-speech-and-time-series>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lee, S., Song, C. H., Park, R. S., Park, M. E., Han, K. M., Kim, J., Choi, M., Ghim, Y. S., & Woo, J.-H. (2016). GIST-PM-Asia v1: Development of a numerical system to improve particulate matter forecasts in South Korea using geostationary satellite-retrieved aerosol optical data over Northeast Asia. *Geoscientific Model Development*, 9(1), 17–39. <https://doi.org/10.5194/gmd-9-17-2016>
- Li, Y., Xie, W., & Li, H. (2017). Hyperspectral image reconstruction by deep convolutional neural network for classification. *Pattern Recognition*, 63, 371–383. <https://doi.org/10.1016/j.patcog.2016.10.019>
- Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., & Catanzaro, B. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. In V. Ferrari, M. Hebert, C.

- Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11215, pp. 89–105). Springer International Publishing. https://doi.org/10.1007/978-3-030-01252-6_6
- Lops, Y., Choi, Y., Eslami, E., & Sayeed, A. (2019). Real-time 7-day forecast of pollen counts using a deep convolutional neural network. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-019-04665-0>
- Ma, Z., Dey, S., Christopher, S., Liu, R., Bi, J., Balyan, P., & Liu, Y. (2021). A review of statistical methods used for developing large-scale and long-term PM2.5 models from satellite data. *Remote Sensing of Environment*, 112827. <https://doi.org/10.1016/j.rse.2021.112827>
- Malm, W. C., Schichtel, B. A., & Pitchford, M. L. (2011). Uncertainties in PM2.5 Gravimetric and Speciation Measurements and What We Can Learn from Them. *Journal of the Air & Waste Management Association*, 61(11), 1131–1149. <https://doi.org/10.1080/10473289.2011.603998>
- Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A., & Cahill, T. A. (1994). Spatial and seasonal trends in particle concentration and optical extinction in the United States. *Journal of Geophysical Research*, 99(D1), 1347. <https://doi.org/10.1029/93JD02916>
- Martin, R. V. (2008). Satellite remote sensing of surface air quality. *Atmospheric Environment*, 42(34), 7823–7843. <https://doi.org/10.1016/j.atmosenv.2008.07.018>
- Mhawish, A., Kumar, M., Mishra, A. K., Srivastava, P. K., & Banerjee, T. (2018). Chapter 3 - Remote Sensing of Aerosols From Space: Retrieval of Properties and Applications. In T. Islam, Y. Hu, A. Kokhanovsky, & J. Wang (Eds.), *Remote Sensing of Aerosols, Clouds, and Precipitation* (pp. 45–83). Elsevier. <https://doi.org/10.1016/B978-0-12-810437-8.00003-7>
- Mikolov, T., Kombrink, S., Burget, L., Cernocky, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5528–5531. <https://doi.org/10.1109/ICASSP.2011.5947611>
- Molod, A., Takacs, L., Suarez, M., & Bacmeister, J. (2015). Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA2. *Geoscientific Model Development*, 8(5), 1339–1356. <https://doi.org/10.5194/gmd-8-1339-2015>
- Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, 114(4), 358–371. <https://doi.org/10.1016/j.biosystemseng.2012.08.009>
- Myllyvirta, L. (2020). Quantifying the Economic Costs of Air Pollution from Fossil Fuels. *Centre for Research on Energy and Clean Air*. <https://energyandcleanair.org/publications/costs-of-air-pollution-from-fossil-fuels/>

- Pandis, S. N. (Ed.). (2004). *Atmospheric aerosol processes*. Cambridge University Press.
- Pouyaei, A., Choi, Y., Jung, J., Sadeghi, B., & Song, C. H. (2020). Concentration Trajectory Route of Air pollution with an Integrated Lagrangian model (C-TRAIL model v1.0) derived from the Community Multiscale Air Quality Modeling (CMAQ model v5.2). *Geoscientific Model Development Discussions*, 1–30. <https://doi.org/10.5194/gmd-2019-366>
- Randles, C. A., da Silva, A. M., Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., & Flynn, C. J. (2017). The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System Description and Data Assimilation Evaluation. *Journal of Climate*, 30(17), 6823–6850. <https://doi.org/10.1175/JCLI-D-16-0609.1>
- Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., ... Woollen, J. (2011). MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *Journal of Climate*, 24(14), 3624–3648. <https://doi.org/10.1175/JCLI-D-11-00015.1>
- Rifkin, R. M., & Lippert, R. A. (n.d.). *Notes on Regularized Least-Squares*. 10.
- Rulloni, V., Bustos, O., & Flesia, A. G. (2012). Large gap imputation in remote sensed imagery of the environment. *Computational Statistics & Data Analysis*, 56(8), 2388–2403. <https://doi.org/10.1016/j.csda.2012.02.022>
- Sayeed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Salman, A. K., Lee, J.-B., Park, H.-J., & Choi, M.-H. (2021). A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Scientific Reports*, 11(1), 1–8. <https://doi.org/10.1038/s41598-021-90446-6>
- Sayeed, A., Lops, Y., Choi, Y., Jung, J., & Salman, A. K. (2021). Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks. *Atmospheric Environment*, 253, 118376. <https://doi.org/10.1016/j.atmosenv.2021.118376>
- Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons.
- Storey, J., Scaramuzza, P., Schmidt, G., & Barsi, J. (2005). *LANDSAT 7 SCAN LINE CORRECTOR-OFF GAP-FILLED PRODUCT DEVELOPMENT*. 13.
- US EPA, O. (2015, July 8). *Benefits and Costs of the Clean Air Act 1990-2020, the Second Prospective Study* [Overviews and Factsheets]. <https://www.epa.gov/clean-air-act-overview/benefits-and-costs-clean-air-act-1990-2020-second-prospective-study>
- US EPA, O. (2020, September 2). *Regional and Geographic Offices* [Collections and Lists]. <https://www.epa.gov/aboutepa/regional-and-geographic-offices>

- Vu, B. N., Bi, J., Wang, W., Huff, A., Kondragunta, S., & Liu, Y. (2022). Application of geostationary satellite and high-resolution meteorology data in estimating hourly PM_{2.5} levels during the Camp Fire episode in California. *Remote Sensing of Environment*, 271, 112890. <https://doi.org/10.1016/j.rse.2022.112890>
- World Health Organization. (2018). *Global status report on road safety 2018*. World Health Organization. <https://apps.who.int/iris/handle/10665/276462>
- Wu, W.-S., Purser, R. J., & Parrish, D. F. (2002). Three-Dimensional Variational Analysis with Spatially Inhomogeneous Covariances. *Monthly Weather Review*, 130(12), 2905–2916. [https://doi.org/10.1175/1520-0493\(2002\)130<2905:TDVAWS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2)
- Wu, Y., Arapi, A., Huang, J., Gross, B., & Moshary, F. (2018). Intra-continental wildfire smoke transport and impact on local air quality observed by ground-based and satellite remote sensing in New York City. *Atmospheric Environment*, 187, 266–281. <https://doi.org/10.1016/j.atmosenv.2018.06.006>
- Yu, C., Chen, L., Su, L., Fan, M., & Li, S. (2011). Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth. *2011 19th International Conference on Geoinformatics*, 1–6. <https://doi.org/10.1109/GeoInformatics.2011.5981052>
- Zhang, C., Li, W., & Travis, D. (2007). Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach. *International Journal of Remote Sensing*, 28(22), 5103–5122. <https://doi.org/10.1080/01431160701250416>
- Zhang, C.-Y., Chen, C. L. P., Gan, M., & Chen, L. (2015). Predictive Deep Boltzmann Machine for Multiperiod Wind Speed Forecasting. *IEEE Transactions on Sustainable Energy*, 6(4), 1416–1425. <https://doi.org/10.1109/TSTE.2015.2434387>
- Zhang, Q., Geng, G., Wang, S., Richter, A., & He, K. (2012). Satellite remote sensing of changes in NO_x emissions over China during 1996–2010. *Chinese Science Bulletin*, 57(22), 2857–2864. <https://doi.org/10.1007/s11434-012-5015-4>
- Zhang, Q., Yuan, Q., Zeng, C., Li, X., & Wei, Y. (2018). Missing Data Reconstruction in Remote Sensing Image With a Unified Spatial–Temporal–Spectral Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8), 4274–4288. <https://doi.org/10.1109/TGRS.2018.2810208>