atmosphere

MDPI

Article

Estimation of $PM_{2.5}$ Concentrations in New York State: Understanding the Influence of Vertical Mixing on Surface $PM_{2.5}$ Using Machine Learning

Wei-Ting Hung [1,*], Cheng-Hsuan (Sarah) Lu [1,2,*], Stefano Alessandrini [3], Rajesh Kumar [3] and Chin-An Lin [1]

1 Atmospheric Sciences Research Center, University at Albany, State University of New York, Albany, NY 12222, USA; clin9@albany.edu
2 Joint Center for Satellite Data Assimilation, Boulder, CO 80301, USA
3 Research Applications Laboratory, National Center of Atmospheric Research, Boulder, CO 80305, USA; alessand@ucar.edu (S.A.); rkumar@ucar.edu (R.K.)
* Correspondence: whung@albany.edu (W.-T.H.); clu4@albany.edu (C.-H.L.)

Received: 24 September 2020; Accepted: 30 November 2020; Published: 30 November 2020

check for updates

Abstract: In New York State (NYS), episodic high fine particulate matter ($PM_{2.5}$) concentrations associated with aerosols originated from the Midwest, Mid-Atlantic, and Pacific Northwest states have been reported. In this study, machine learning techniques, including multiple linear regression (MLR) and artificial neural network (ANN), were used to estimate surface $PM_{2.5}$ mass concentrations at air quality monitoring sites in NYS during the summers of 2016–2019. Various predictors were considered, including meteorological, aerosol, and geographic predictors. Vertical predictors, designed as the indicators of vertical mixing and aloft aerosols, were also applied. Overall, the ANN models performed better than the MLR models, and the application of vertical predictors generally improved the accuracy of $PM_{2.5}$ estimation of the ANN models. The leave-one-out cross-validation results showed significant cross-site variations and were able to present the different predictor-$PM_{2.5}$ correlations at the sites with different $PM_{2.5}$ characteristics. In addition, a joint analysis of regression coefficients from the MLR model and variable importance from the ANN model provided insights into the contributions of selected predictors to $PM_{2.5}$ concentrations. The improvements in model performance due to aloft aerosols were relatively minor, probably due to the limited cases of aloft aerosols in current datasets.

Keywords: air quality; $PM_{2.5}$; machine learning; vertical mixing

1. Introduction

Fine particulate matter ($PM_{2.5}$) (particulate matter with aerodynamic diameter less than 2.5 μm) is one of the criteria of air pollutants because of its detrimental impacts on human health and the environment [1,2]. Previous studies have reported that the exposure to high $PM_{2.5}$ concentrations can increase the risk of respiratory diseases and mortality [3,4]. Many processes can affect ground-level $PM_{2.5}$ concentrations, including emissions, removal (e.g., deposition), transport, aerosol physical processes (e.g., nucleation), and atmospheric chemistry [5–9]. These processes are potentially affected by meteorological conditions (e.g., surface temperature and horizontal winds) [10–14]. Yu et al. (2008) [9] used the Eta-Community Multiscale Air Quality (CMAQ) coupled model to estimate the surface $PM_{2.5}$ concentrations over the eastern United States (US) during the summer of 2004, and indicated that aerosol physical and chemical processes dominated $PM_{2.5}$ concentration. Dawson et al. (2007) [10] analyzed the sensitivities of $PM_{2.5}$ concentrations to meteorological variables, and showed that temperature, absolute humidity, wind speed, mixing layer height, and precipitation had the strongest effects on $PM_{2.5}$

concentrations. Additionally, atmospheric vertical mixing has been shown to significantly impact $PM_{2.5}$ concentrations [7,15–19]. In Zhang et al. (2020) [19], the radar wind profiler measurements showed weak vertical wind shears at a shallower planetary boundary layer (PBL) under polluted conditions, as evidence of weak vertical mixing leading to strong $PM_{2.5}$ accumulation in the PBL. They indicated that strong vertical wind shears were associated with strong vertical mixing and often accompanied with low $PM_{2.5}$ concentrations. Strong winds above the PBL were also favorable for the transport of aerosols, which could potentially affect surface $PM_{2.5}$ concentrations through vertical mixing.

Various approaches have been applied for surface $PM_{2.5}$ estimation, including chemistry transport models (CTMs) [9,20,21] and land use regression (LUR) models [22,23]. Statistical approaches based on the relationships between satellite aerosol optical depth (AOD) and surface $PM_{2.5}$ concentrations have also been applied [24–26]. Recently, machine learning (ML), an application of artificial intelligence (AI), has become an increasingly popular approach for $PM_{2.5}$ estimation [27–29]. ML also provides insights into contributions of different influencing factors to $PM_{2.5}$ concentrations. For instance, Reid et al. (2015) [29] estimated the $PM_{2.5}$ concentrations during the Northern California wildfires in 2008. Various predictors were used in ML training, including meteorological variables, such as temperature and humidity, land use variables, such as the distance to the nearest emission source, geographic variables, such as site location and Julian date, satellite AOD measurements, and CTM simulated $PM_{2.5}$ concentrations. Meteorological, land use, and geographic variables provided the baseline $PM_{2.5}$ estimation. The application of AOD and CTM products further improved model performance by providing the estimation of total-column aerosol loading and prior knowledge of the aerosol physical processes and chemical reactions, respectively. Furthermore, several studies focused on the influence of AOD on $PM_{2.5}$ estimation and indicated that AOD could improve the accuracy of $PM_{2.5}$ estimation with increased correlation coefficients and decreased model errors [30–33]. However, Yao et al. (2018) [34] reported an unchanged model performance when using AOD in ML training, probably due to complex terrain, uncertainties of cloud filtering and the presence of aloft aerosols.

In New York State (NYS), the $PM_{2.5}$ concentrations have been decreasing continuously over the past decades [35–38]. Rattigan et al. (2015) [37] analyzed the 2000–2014 observed $PM_{2.5}$ concentrations at 16 air quality monitoring sites across the state and reported a downward trend with decreases of 4–7 $\mu g\ m^{-3}$ on annual scale. The annual average $PM_{2.5}$ concentrations at 15 sites were 9–17 $\mu g\ m^{-3}$ and 6–10 $\mu g\ m^{-3}$ in 2000 and 2014, respectively, while the annual average $PM_{2.5}$ concentration at the Whiteface Mountain site decreased from about 6 $\mu g\ m^{-3}$ in 2000 to 4 $\mu g\ m^{-3}$ in 2014. In addition, according to the New York State Ambient Air Quality Report for 2019 (https://www.dec.ny.gov/docs/air_pdf/2019airqualreport.pdf), the 2019 $PM_{2.5}$ annual averages ranged from 5 to 9 $\mu g\ m^{-3}$, except for the Whiteface Mountain site with an annual average around 3 $\mu g\ m^{-3}$.

However, episodic high $PM_{2.5}$ concentrations across NYS have been reported. During wintertime, the high $PM_{2.5}$ concentrations have been attributed to local heating sources and lower mixing layer heights [36,38]. As for summer, the high $PM_{2.5}$ concentrations have been attributed to local emissions [39,40], anthropogenic aerosols transported from the Midwest and the Great Lake region [39–43], and long-range transported smoke aerosols from the western US and Canada [44–48]. Climatically, NYS is affected by the prevailing westerlies. Additionally, the Bermuda High locating over the western Atlantic Ocean introduces southerly and southwesterly winds along the east coast in summertime. Under the influences of synoptic flows, air masses could transport aerosols from the polluted areas to NYS. These aerosols could be potentially transported from free troposphere to the surface, resulting in increased $PM_{2.5}$ concentrations. In Dutkiewicz et al. (2004) [43], 1-year observations of the concentrations of $PM_{2.5}$ sulfate showed that more than 40% of the high sulfate concentrations were associated with westerly flows and around 20–30% were associated with southwesterly flows, reflecting the influences of transported pollutants from the Midwest and Mid-Atlantic states, respectively. The contribution of such transported pollutants was more significant at rural sites, as up to 60% of $PM_{2.5}$ sulfate was associated with transported pollutants on annual basis. On the other hand, a statewide event of high $PM_{2.5}$ concentrations in August 2018 caused by transported smoke aerosols was investigated in Hung et al. (2020) [45]. Multi-platform measurements and model simulations demonstrated the long-range transport of smoke aerosols from the western states to NYS. These smoke aerosols transported in free troposphere before reaching NYS and descended to around 2 km a.g.l. driven by synoptic downward mixings. The PBL entrainment further brought these aloft aerosols to the surface, resulting in a threefold increase (from 8 to 24 μg m^{-3}) in the average $PM_{2.5}$ concentrations.

Since transported aerosols make significant contributions to the high $PM_{2.5}$ concentrations in NYS, understandings of the relationships between the vertical mixing of these aloft aerosols and surface $PM_{2.5}$ concentrations are critical for air quality forecast, monitoring and management. Additionally, studies exploring the roles of vertical mixing in $PM_{2.5}$ estimation are needed. In this study, ML techniques were used to estimate the $PM_{2.5}$ mass concentrations at air quality monitor sites in NYS during summer seasons (July, August and September, JAS) of 2016–2019. Multiple predictors were applied, including meteorological, aerosol and geographic variables. Predictors associated with vertical mixing and aloft aerosols were also considered. The statistical correlations between selected predictors and $PM_{2.5}$ concentrations were investigated. Additionally, to understand the influences of predictors on the $PM_{2.5}$ concentration in NYS, the results at monitoring sites with different ambient conditions were discussed.

2. Experiments

2.1. Datasets

The variables used in this study are summarized in Table 1 and are briefly described in this section. Details can be found in the cited references.

Table 1. List of variables used in this study.

Variable	Source	Level	Spatial Resolution	Temporal Resolution
	Target			
$PM_{2.5}$ observation (μg m^{-3})	EPA AQS [1]	Surface		Hourly
	Meteorological predictors			
Surface pressure (Pa)	HRRR [2]	Surface	3 km	3-hourly
Temperature (K)	HRRR	2 m a.g.l.	3 km	3-hourly
Relative humidity (%)	HRRR	2 m a.g.l.	3 km	3-hourly
U-component of horizontal wind (m s^{-1})	HRRR	10 m a.g.l.	3 km	3-hourly
V-component of horizontal wind (m s^{-1})	HRRR	10 m a.g.l.	3 km	3-hourly
Planetary boundary layer height (m)	HRRR		3 km	3-hourly
	Aerosol predictors			
Aerosol optical depth	VIIRS [3]	Total column	0.25° × 0.25°	Daily
$PM_{2.5}$ concentration (μg m^{-3})	MERRA-2 [4]	Surface	0.5° × 0.625°	Hourly
	Geographic predictors			
Latitude	EPA AQS			
Longitude	EPA AQS			
Altitude (m)	EPA AQS			
Vegetation index	VIIRS [5]	Surface	0.05° × 0.05°	Monthly
Weekday				Daily
	Vertical predictors			
Wind shear (s^{-1})	HRRR	Surface—850 hPa 850—700 hPa 700—500 hPa	3 km	3-hourly
Average vertical velocity (Pa s^{-1})	HRRR	Surface—500 hPa	3 km	3-hourly
Ratio of AOD change rate to PM change rate	VIIRS		0.25° × 0.25°	Daily
	EPA AQS			Daily

Note: [1] https://aqs.epa.gov/aqsweb/documents/data_api.html;
[2] http://home.chpc.utah.edu/~u0553130/Brian_Blaylock/cgi-bin/hrrr_download.cgi;
[3] https://www.star.nesdis.noaa.gov/smcd/emb/viirs_aerosol/products_gridded.php;
[4] https://disc.gsfc.nasa.gov/datasets/M2T1NXAER_5.12.4/summary;
[5] https://lpdaac.usgs.gov/products/vnp13c2v001/.

2.1.1. Surface $PM_{2.5}$ Observations

The US Environmental Protection Agency (EPA) collects real-time air quality measurements from over 2000 surface monitoring sites nationwide maintained by state or local air quality agencies. In NYS, air quality data are collected and quality controlled by the NYS Department of Environmental Conservation (NYSDEC). In this study, hourly $PM_{2.5}$ mass concentrations from 21 monitoring sites across NYS were used (Figure 1; Table 2). According to the US EPA and Squizzato et al. (2018) [38], these sites consisted of 18 urban/suburban sites and 3 rural sites. The urban/suburban sites were divided into New York City (NYC) metropolitan sites and upstate NY (UNY) sites based on their locations defined by the US EPA core based statistical areas (CBSA). As a result, in this study, 21 selected sites were categorized into: (1) 5 UNY sites, which is a group of sites in Buffalo, Rochester, and Albany areas, (2) 3 rural sites, and (3) 13 NYC sites, which located in the New York–Newark–Jersey City area. The $PM_{2.5}$ concentration measurements at three sites (marked with asterisks in Table 2) are based on the Federal Equivalence Method (FEM), while others are based on the Tapered Element Oscillating Microbalances (TEOM) technology.

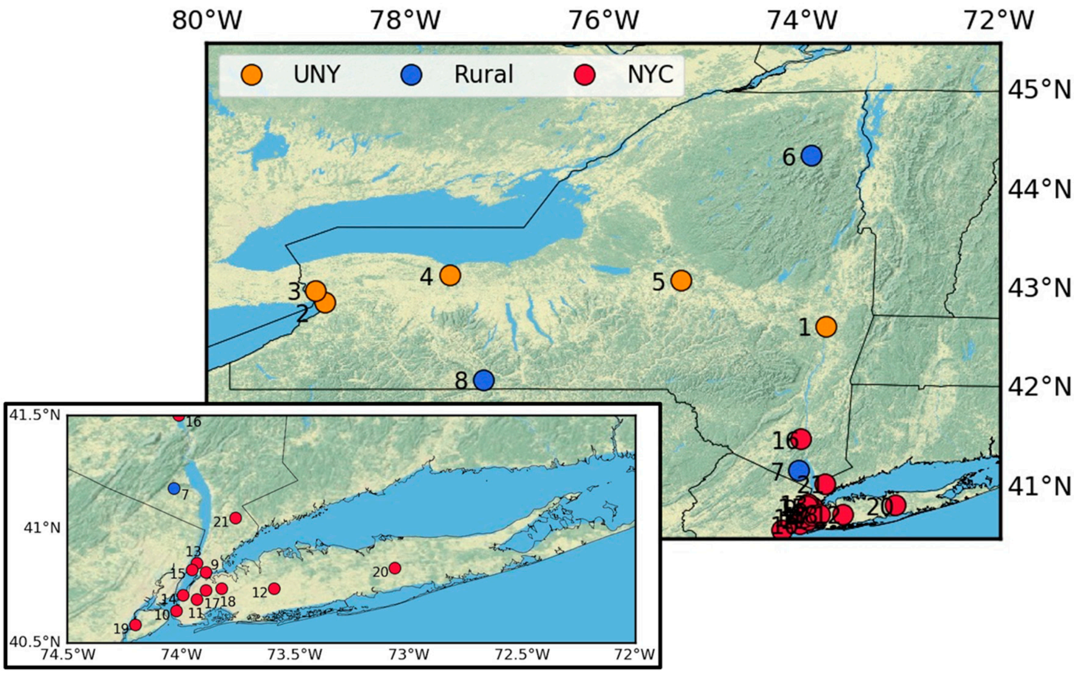

Figure 1. Topographic map with the $PM_{2.5}$ monitoring sites used in this study. Orange, blue and red circles are UNY, rural and NYC sites, respectively. Site labels are referred from Table 2.

Table 2. List of $PM_{2.5}$ monitor sites in NYS used in this study. Site labels are referred to in Figure 1. Sites using Federal Equivalence Method (FEM) are marked with asterisks *. Sites 1–5, 6–8 and 9–21 are UNY, rural and NYC sites, respectively.

Label	Name	ID Number	Latitude	Longitude	Altitude (m)	Type
1	Albany	360010005	42.64	−73.75	7	UNY
2	Buffalo	360290005	42.88	−78.81	185	UNY
3	Tonawanda II	360291014	43	−78.9	182	UNY
4	Rochester *	360551007	43.15	−77.55	137	UNY
5	Utica	360652001	43.1	−75.22	139	UNY
6	Whiteface Mountain	360310003	44.36	−73.9	599	Rural
7	Rockland County	360870005	41.18	−74.03	140	Rural
8	Pinnacle State Park *	361010003	42.1	−77.21	507	Rural

Table 2. *Cont.*

Label	Name	ID Number	Latitude	Longitude	Altitude (m)	Type
9	Bronx	360050112	40.81	−73.89	20	NYC
10	PS 314	360470052	40.64	−74.02	26	NYC
11	PS 274	360470118	40.69	−73.93	18	NYC
12	Esienhower Park	360590005	40.74	−73.59	27	NYC
13	IS 143	360610115	40.85	−73.93	0	NYC
14	Division St.	360610134	40.71	−73.99	17	NYC
15	CCNY	360610135	40.82	−73.95	45	NYC
16	Newburgh	360710002	41.5	−74.01	127	NYC
17	Maspeth	360810120	40.73	−73.89	31	NYC
18	Queens *	360810124	40.74	−73.82	25	NYC
19	FKILL	360850111	40.58	−74.2	3	NYC
20	Holtsville	361030009	40.83	−73.06	45	NYC
21	White Plain	361192004	41.05	−73.76	64	NYC

2.1.2. Meteorological Predictors

Several meteorological predictors were obtained in this study, including 2 m temperature (T), 2 m relative humidity (RH), surface pressure (PS), planetary boundary layer height (PBLH), and the u and v components of 10 m horizontal winds (U, V). These variables were taken from the analysis fields of the High-Resolution Rapid Refresh (HRRR) [49], which is an atmospheric model developed by the NOAA/Earth System Research Laboratories (ESRL)/Global Systems Laboratory (GSL). HRRR has 3 km horizontal resolution and 51 vertical levels in hybrid coordinates, and provides hourly analysis over the contiguous US (CONUS) and Alaska. Details about HRRR can be found at https://rapidrefresh.noaa.gov/hrrr/ and http://www.nco.ncep.noaa.gov/pmb/products/hrrr/.

2.1.3. Aerosol Predictors

AOD measurements from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor onboard the Suomi National Polar-orbiting Partnership (S-NPP) satellite, launched in October 2011, were used. VIIRS measures 22 spectrum channels in the range of 412–12,050 nm, including imagery bands, moderate resolution bands (M-bands) and the day–night band. The wide spectral range allows VIIRS to provide multiple land and atmosphere products and the M-bands are mainly used for aerosol retrieval [50]. VIIRS provides daily global coverage with 750 m resolution and only daytime measurements are used for AOD retrieval. In this study, level 3 Environmental Data Record (EDR) daily gridded AOD at 550 nm products were used.

In addition, surface $PM_{2.5}$ mass concentrations from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) [51] were used. MERRA-2 is a global atmospheric reanalysis developed by the National Aeronautics and Space Administration's (NASA's) Global Modeling and Assimilation Office (GMAO). MERRA-2 uses the Goddard Earth Observing System, Version 5 (GEOS-5) Atmospheric General Circulation Model (AGCM) coupled with the Goddard Chemistry Aerosol Radiation and Transport (GOCART) model [52]. Aerosol and meteorological observations are jointly assimilated in MERRA-2. MERRA-2 aerosol reanalysis considers aerosol emissions, transport, removal processes, and chemistry. Details can be found in Buchard et al. (2017) [53] and Randles et al. (2017) [54]. A brief comparison between HRRR and MERRA-2 meteorological fields (T, RH, PS, U and V) showed that two models are in good agreement with correlation coefficients higher than 0.7 (Appendix A).

2.1.4. Geographic Predictors

In this study, the monthly enhanced vegetation index (VI) [55,56] products from VIIRS S-NPP were used as terrestrial vegetation estimation. Weekday and geographic information (latitude, longitude and altitude) of selected monitor sites, referred from EPA AQS database, were also used.

2.1.5. Vertical Predictors

To describe the atmospheric vertical mixing, vertical wind shears (VWS) of three layers, including surface—850 hPa (Low-Level; L-VWS), 850–700 hPa (Mid-Level; M-VWS) and 700–500 hPa (High-Level; H-VWS) from HRRR analysis, were used in this study. VWS were defined as the gradients of horizontal winds between the top and bottom model levels. Note that this study only considered the magnitude of VWS, which was computed from the u and v components of VWS. The average vertical velocities of the layer of surface to 500 hPa (W_avg) were also used.

Furthermore, to indicate the presence of aloft aerosols, the daily change rates (R) of *AOD* and $PM_{2.5}$ concentrations at monitor sites were calculated as follows:

$$R_{AOD,\ d} = \frac{AOD_d - AOD_{d-1}}{AOD_{d-1}} \tag{1}$$

$$R_{PM2.5,\ d} = \frac{PM2.5_d - PM2.5_{d-1}}{PM2.5_{d-1}} \tag{2}$$

where d is date. The day-by-day variation of AP_ratio reflects the change in aerosol vertical distribution. Since *AOD* and $PM_{2.5}$ concentrations present the aerosol loadings in the total-column atmosphere and near the surface, respectively, their daily change rates should be comparable if most of the aerosols are near the surface. R_{AOD} should be higher than $R_{PM2.5}$ if there are aerosols aloft. In contrast, under weak advection with no change in aloft aerosols, $R_{PM2.5}$ could be higher than R_{AOD}, since surface $PM_{2.5}$ concentration is mainly determined by local emissions. In this study, the ratio of R_{AOD} to $R_{PM2.5}$ (AP_ratio) was used as the indicator of aloft aerosols.

2.1.6. Data Processing

Datasets in the domain of 40.5°N –45.5° N, 72° W–80° W during the summer seasons (JAS) of 2016–2019, 368 days in total, were used. Four data processing steps were taken prior to training. First, spatial linear interpolations were applied for MERRA-2 data, and 3×3 grid averages were computed for satellite AOD and VI as the representative at the location of selected air quality sites. Second, the averages of daytime data during 0700–1800 LT (1200–2300 UTC), except for AOD and VI, were calculated. Afterward, days with missing data were removed. The outliers (i.e., values beyond average ±3 standard deviations) of $PM_{2.5}$ observation and AOD were also removed to exclude extreme cases (about 1.5% of the data). Lastly, aerosol predictors were transformed to become Gaussian distribution by taking square roots on $PM_{2.5}$ observation and AOD and taking a log of MERRA-2 $PM_{2.5}$ concentration.

The assumption of independence of predictors, required for ML algorithms, is examined by correlation matrix (Appendix B). While the assumption is not strictly met, the impact on training results may not be significant given only one pair exceeding 0.8. It is worth mentioning that, although some variables may be connected physically (e.g., T and PBLH), the correlations between these variables are relatively weak.

2.2. Model Configuration

Two ML algorithms were used in this study: multiple linear regression (MLR) and artificial neural network (ANN). MLR is a statistical technique which estimates the relationships between several explanatory variables (i.e., predictors) and a response variable (i.e., target) by fitting a linear regression to the ground truth ($PM_{2.5}$ observations in this study). It generates a linear regression of given variables, including the coefficients for predictors and intercept. In this study, the Python Scikit-Learn package [57] was used to build the MLR models.

An artificial neural network (ANN), one of the most popular ML algorithms, is effective for handling complex nonlinear problems, particularly in classification and prediction [58,59]. An ANN model consists of a set of layers, including the input layer, multiple hidden layers, and the output layer. There is no specific rule to decide how many hidden layers an ANN model should have but one hidden layer with abundant neurons usually provides good results [58]. In addition, previous studies showed that ANN models performed better with the number of neurons in a range of $(2\sqrt{n} + \varphi, 2n + 1)$, where n and φ are the numbers of predictors and targets, respectively [60]. In this study, the Python Keras and Tensorflow packages [61] were used. Back-propagation models with one hidden layer obtaining 20 neurons were applied. The maximum number of iterations was set to 1000 and an EarlyStopping function (https://keras.io/api/callbacks/early_stopping/) was used to avoid overfitting. Variables were randomly split into the training set (80%) and validation set (20%) in the training process, and an additional testing set, which will be described in the following section, was applied for model evaluation.

To investigate the influence of vertical mixing and aloft aerosols on surface $PM_{2.5}$ concentration, two sets of predictors were applied. Set 1 contained meteorological, aerosol, and geographic predictors, while set 2 contained the same predictors as set 1 but also included vertical predictors. Therefore, four models were applied in this study:

- MLR model with set 1 predictors (MLR-1);
- MLR model and set 2 predictors (MLR-2);
- ANN model with set 1 predictors (ANN-1);
- ANN model with set 2 predictors (ANN-2).

2.3. Statistical Analysis

In this study, leave-one-out cross validation (LOOCV) [62] was used for model performance evaluation. All four models were trained on data from all monitor sites but one and were tested on the leave-out site. This process was repeated until all 21 sites served as the test site once. Therefore, LOOCV ensured independent evaluation of the trained model via comparison of the predicted and observed values at the locations that did not participate in the training. The involvement of spatial dependence allowed LOOCV to provide more reliable estimations of model performance [63]. Three statistical scores were used for model evaluation, including mean bias (MB), coefficient of determination (R^2), and root mean square error (RMSE). These scores were calculated as follows:

$$MB = \frac{1}{n}\sum_{i}^{n} Y_i - X_i \tag{3}$$

$$R^2 = \left(\frac{\Sigma_i^n\left(Y_i - \overline{Y}\right)\Sigma_i^n\left(X_i - \overline{X}\right)}{\sqrt{\Sigma_i^n\left(Y_i - \overline{Y}\right)^2}\sqrt{\Sigma_i^n\left(X_i - \overline{X}\right)^2}}\right)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(Y_i - X_i)^2} \tag{5}$$

where n is number of data points, X is observation, Y is prediction, and $\overline{X}$ and $\overline{Y}$ are the averages of observation and prediction, respectively. The scores at test sites were then averaged as the estimated model errors. Although LOOCV is beneficial for estimating the spatial dependence within trained models, the imbalanced site distribution in NYS (over 60% of the sites are in NYC region) may introduce representativeness issues in the model performance.

To understand the relationships between predictors and $PM_{2.5}$ concentrations, the regression coefficients from the MLR models and permutation importance estimated from the ANN models were analyzed. The values of regression coefficients explained the change in target when applying one unit of change in predictors, and the signs explained the direction of such a change. Permutation importance is an approach for ranking variable importance [64,65]. The goal of this approach is to estimate how model performance changes when breaking the correlations between predictors and target. This is done by randomly shuffling one predictor in the test data, and comparing the statistical scores of the shuffled model and unshuffled model. Larger percentage errors indicate higher importance. This process is repeated until all predictors have been shuffled once. In this study, the percentage error of the RMSE values of two models was regarded as the estimation of variable importance and is calculated as follows.

$$Variable\ importance = \frac{RMSE_{shuffled} - RMSE_{unshuffled}}{RMSE_{unshuffled}} \times 100\% \tag{6}$$

3. Results and Discussion

3.1. Model Performance

Figure 2 illustrates the LOOCV testing results at selected monitoring sites of the four models. The averages of statistical scores are shown in Table 3, and the statistical scores at individual sites are shown in Appendices C and D. Overall, the ANN models performed better than the MLR models with higher R^2 and lower RMSE, and the ANN models showed larger cross-site variations compared to the MLR models. The absolute values of averaged MB of the ANN models (0.63 and 0.29 μg m^{-3}) were higher than the MLR models (0.03 and 0.06 μg m^{-3}). However, the MBs at individual sites of the four models had a similar range of 2 μg m^{-3}, except for the MLR-2 model which had a wider range of ±3 μg m^{-3}. In addition, the near-zero averaged MBs compared to RMSEs indicated that there were negative and positive biases and those non-systemic biases cancelled out in averaging for the MLR and ANN models.

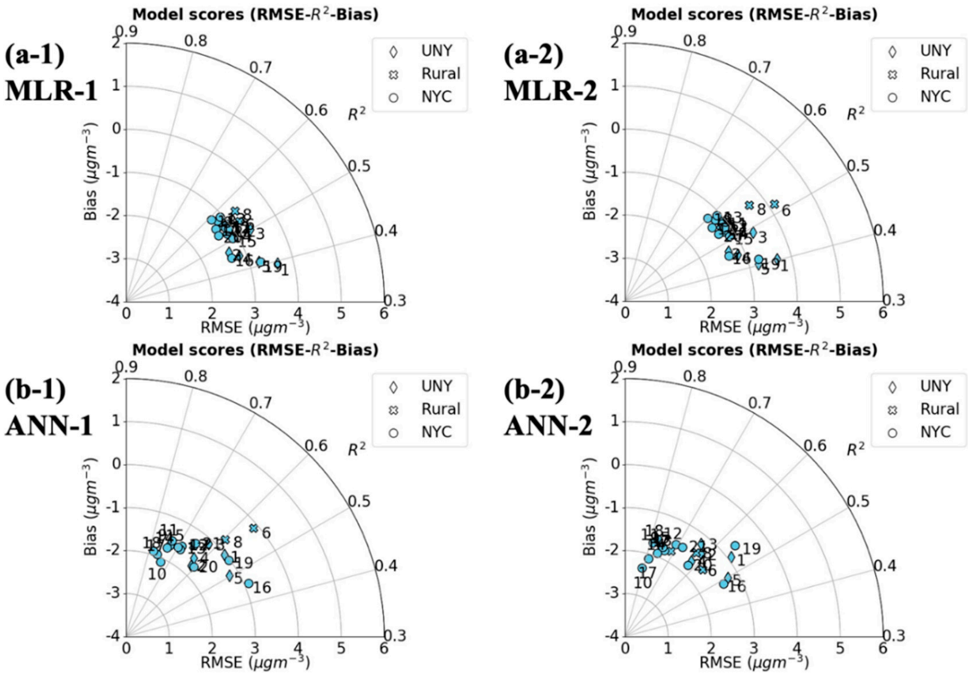

Figure 2. Statistical scores of the testing results of the (**a-1**,**a-2**) MLR and (**b-1**,**b-2**) ANN models with (1) set 1 and (2) set 2 predictors. Site labels are referred from Table 2. Diamonds, crosses and circles are UNY, rural and NYC sites, respectively.

For the MLR models, the application of vertical predictors introduced a neutral impact to the model performance since the differences of averaged statistical scores between MLR-1 and MLR-2 were relatively minor. The RMSE at two rural sites (sites 6 and 8) even increased by 0.5–1 μg m^{-3} when applying vertical predictors (Figure 2a), resulting in a slightly increased averaged RMSE (Table 3).

In contrast, the improvement due to vertical predictors was more significant for the ANN models (Figure 2b). The application of vertical predictors slightly increased the averaged MB for the MLR models (from 0.03 to 0.06 μg m^{-3}), while it reduced the averaged MB for ANN models (from −0.63 to −0.29 μg m^{-3}) (Table 3). Additionally, the range of RMSE changed from (1.92 to 3.89 μg m^{-3}) for the ANN-1 model to (1.64 to 3.32 μg m^{-3}) for the ANN-2 model (Appendix D), showing a general improvement in model performance. It is worth mentioning that the MB and RMSE at sites 6 and 8 decreased by around 2 μg m^{-3}, showing contrary results to the MLR models.

Table 3. Averaged statistical scores of four models at 21 selected sites.

Model	Bias (μg m^{-3})	R-Squared	RMSE (μg m^{-3})
MLR-1	0.03 ± 1.14	0.51 ± 0.06	2.96 ± 0.26
MLR-2	0.06 ± 1.31	0.52 ± 0.06	3.01 ± 0.39
ANN-1	−0.63 ± 0.99	0.65 ± 0.09	2.59 ± 0.45
ANN-2	−0.29 ± 0.88	0.67 ± 0.10	2.42 ± 0.37

3.2. The Site-Variations of Model Performance

The model performance with the influence of vertical predictors on $PM_{2.5}$ prediction at different category of sites were investigated. Figure 3 demonstrates the differences in statistical scores between the models using set 1 and set 2 predictors at selected air quality sites. Table 4 shows the averages of statistic scores of each category of sites. The statistical scores at each site are listed in Appendices C and D. Overall, the results showed variations in model performance across the state, reflecting the different air quality characteristics.

For rural sites, the RMSEs of the MLR-2 model at site 6 and 8 increased from 3.22 and 3.29 μg m^{-3} to 4.13 and 3.65 μg m^{-3}, respectively, compared to the MLR-1 model. The MBs at the two sites also increased, showing that the model performance degraded when applying vertical predictors to the MLR model. In contrast, the performance of the ANN models at three rural sites showed significant improvement with increased R^2 and decreased MB and RMSE. The contrasting performance of the MLR and ANN models could be attributed to the complexity and nonlinearity of the influences of vertical predictors on $PM_{2.5}$ concentrations, which could not be learned by the MLR models. Additional predictors could even reduce the significance of the correlations between other predictors and $PM_{2.5}$ concentrations in the MLR models. It is worth noting that the testing results at site 6 (Whiteface Mountain site, the fifth highest mountain in NYS) showed the most significant differences after applying vertical predictors for both the MLR and ANN models. This is probably because the $PM_{2.5}$ concentrations at Whiteface Mountain are mainly affected by meteorological conditions and transported aerosols [43].

For NYC sites, both the averaged values (Table 4) and testing results at individual site (Appendix C) showed comparable statistical scores for two MLR models. Although the performance of the ANN models showed neutral to positive impacts when applying vertical predictors, the differences in statistical scores were less significant than those at rural sites. This is probably because the air quality of NYC sites is influenced by local anthropogenic emissions and photochemical reactions near the surface, and thus the influence of vertical mixing is limited.

For UNY sites, the statistical scores of the MLR models were comparable, showing limited influence of vertical predictors. As for the ANN models, model performance showed degradation with increased MB at most of the sites after applying vertical predictors (Appendix D). The reductions in RMSE at three sites were relatively minor compared to the increases at two sites, resulting in an increased average value. This may be due to the spatial variability among UNY sites. Unlike NYC sites, UNY sites are a group of urban/suburban sites across the state influenced by different local emissions. The influences of vertical mixing varied among these sites, leading to the degraded testing results on average.

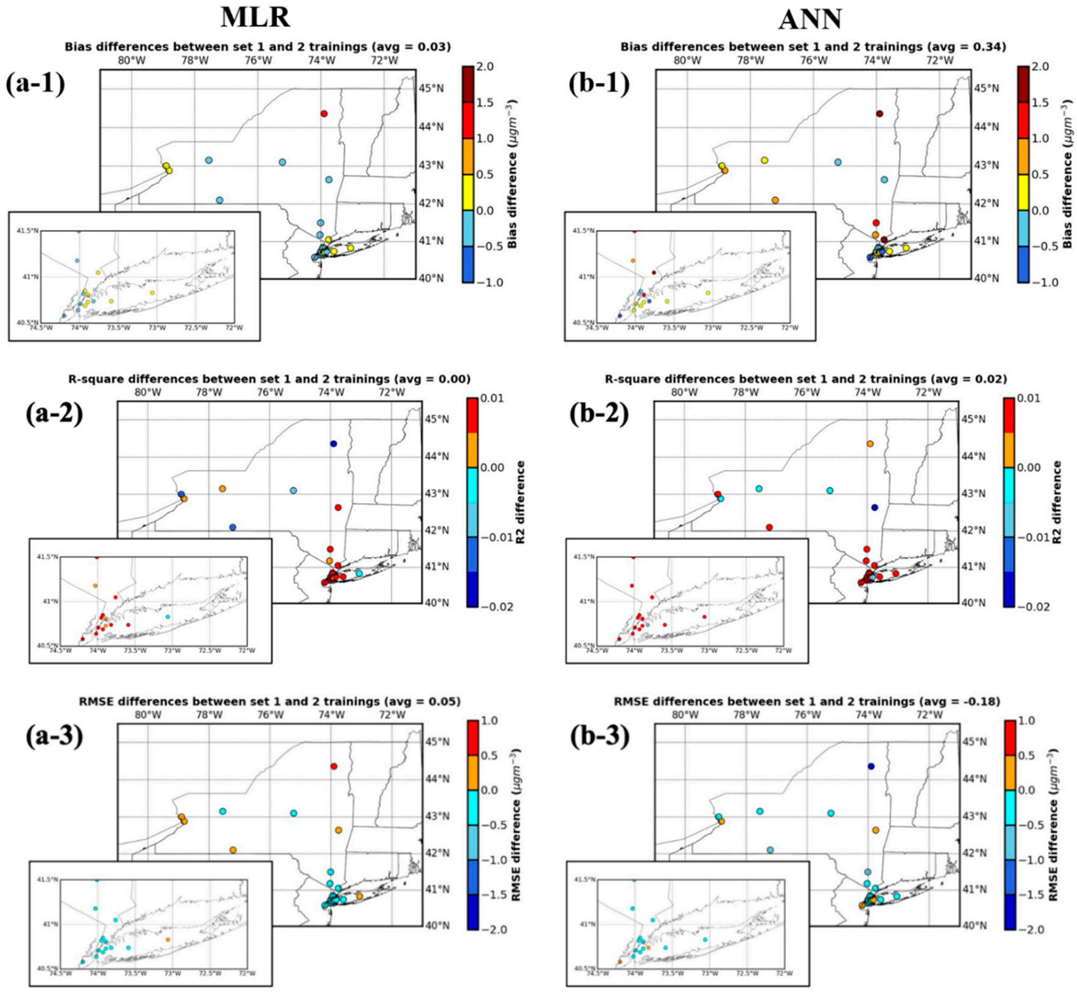

Figure 3. Differences of bias (1), R-squared (2) and RMSE (3) values at selected sites between the (**a-1**,**a-2**,**a-3**) MLR-1 and MLR-2 models, and the (**b-1**,**b-2**,**b-3**) ANN-1 and ANN-2 models. Differences were defined as the statistical scores of set 2 model minus those of set 1 model.

Table 4. Averaged statistical scores of four models at rural, NYC and UNY sites.

Model	Bias (μg m^{-3})	R-Squared	RMSE (μg m^{-3})
	Rural sites		
MLR-1	−0.15 ± 2.04	0.55 ± 0.01	3.11 ± 0.20
MLR-2	0.07 ± 2.67	0.54 ± 0.02	3.53 ± 0.54
ANN-1	−2.10 ± 1.00	0.61 ± 0.04	3.25 ± 0.51
ANN-2	−1.02 ± 0.70	0.64 ± 0.06	2.40 ± 0.12
	NYC sites		
MLR-1	0.09 ± 0.80	0.53 ± 0.05	2.85 ± 0.16
MLR-2	0.10 ± 0.80	0.54 ± 0.05	2.84 ± 0.15
ANN-1	−0.42 ± 0.76	0.68 ± 0.09	2.42 ± 0.33
ANN-2	−0.19 ± 0.89	0.71 ± 0.09	2.31 ± 0.38
	UNY sites		
MLR-1	−0.04 ± 1.09	0.45 ± 0.04	3.13 ± 0.35
MLR-2	−0.05 ± 1.12	0.44 ± 0.04	3.15 ± 0.36
ANN-1	−0.31 ± 0.70	0.59 ± 0.05	2.66 ± 0.29
ANN-2	−0.13 ± 0.76	0.58 ± 0.05	2.71 ± 0.24

Additionally, to better understand the influence of vertical predictors on $PM_{2.5}$ prediction at each category of sites, the testing results at three sites with the lowest RMSEs were analyzed. The PS 314, Rockland County, and Rochester sites were selected as the representatives of NYC, rural and UNY sites, respectively. Additionally, since vertical predictors showed limited contributions in the MLR models in previous discussions, only the results from the ANN models were discussed. Figure 4 illustrates the scatter plots between observed and predicted $PM_{2.5}$ concentrations from the ANN models and the corresponding data plots at selected sites. For PS 314 site (Figure 4a), data plots showed that the ANN-2 model had better performance in predicting spikes with smaller differences between observations and predictions, compared to the ANN-1 model. Although the differences were small, the positive effects of vertical predictors on $PM_{2.5}$ concentrations at NYC sites were not neglectable. Similarly, the testing results of the ANN models at Rockland County site (Figure 4b) showed that the ANN-2 model had better capability of predicting spikes. Although the ANN-2 model still showed outliers in the scatter plot, it provided more accurate estimations with higher R^2 and lower RMSE. On the other hand, the testing results from two ANN models at Rochester site (Figure 4c) were comparable, showing a limited influence of vertical mixing.

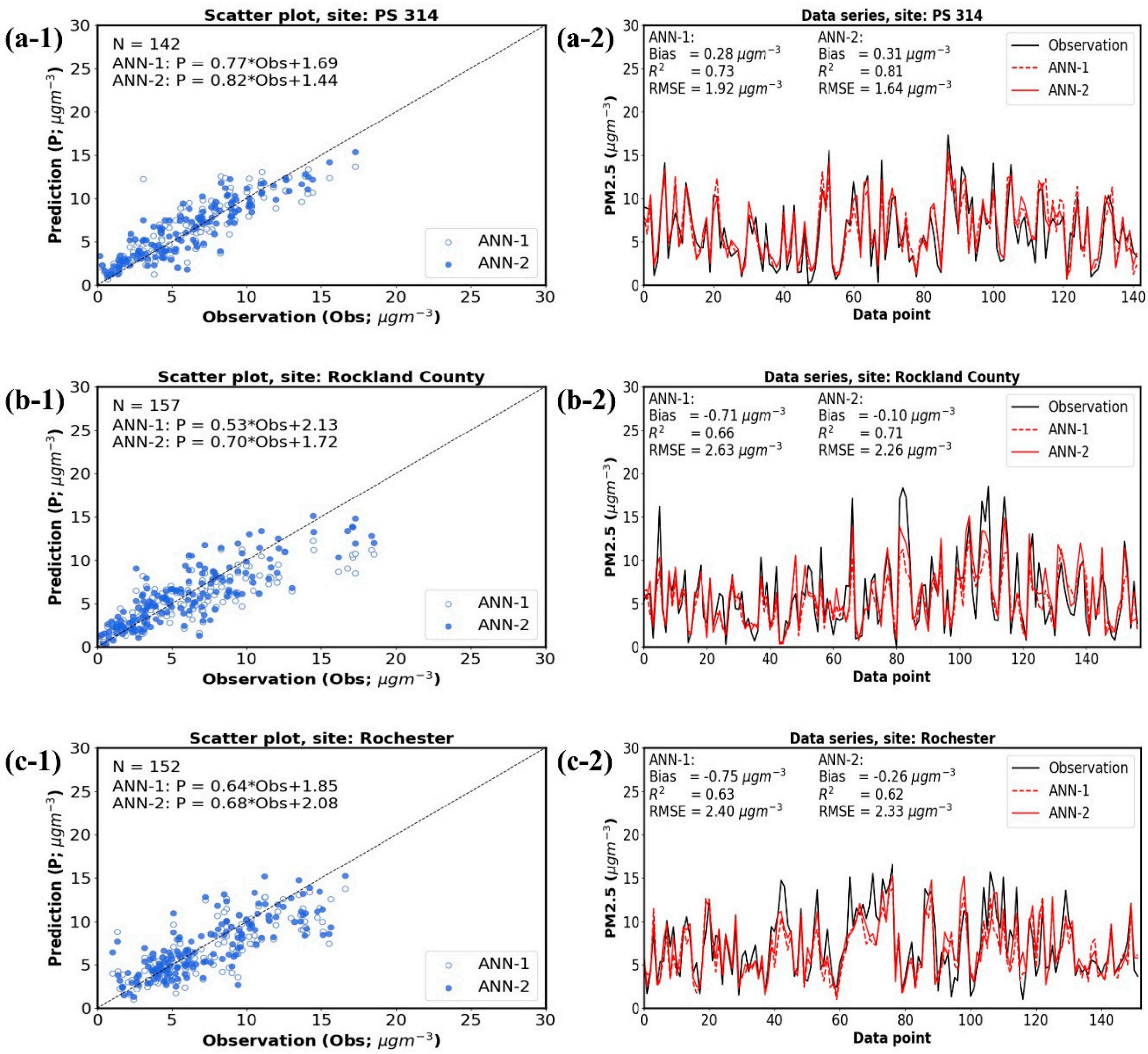

Figure 4. Scatter plots, with $PM_{2.5}$ observation as x-axis and prediction as y-axis, and data plots, with available data point as x-axis and $PM_{2.5}$ concentration as y-axis, of the testing results of the ANN models at (**a-1**,**a-2**) PS 314, (**b-1**,**b-2**) Rockland County and (**c-1**,**c-2**) Rochester sites. The data point in data plots are composited from four summers and each point represents one day.

3.3. The Contributions of Predictors to Surface $PM_{2.5}$ Concentrations

In this section, the regression coefficients generated by the MLR models and variable importance estimated from the ANN models were investigated. Figures 5–7 show the signs of regression coefficients and variable importance at the PS 314, Rochester, and Rockland County sites, respectively. Note that Lat, Lon, and Alt are fixed values at each site, thus shuffling them did not affect the results. The values of regression coefficients are listed in Appendix E.

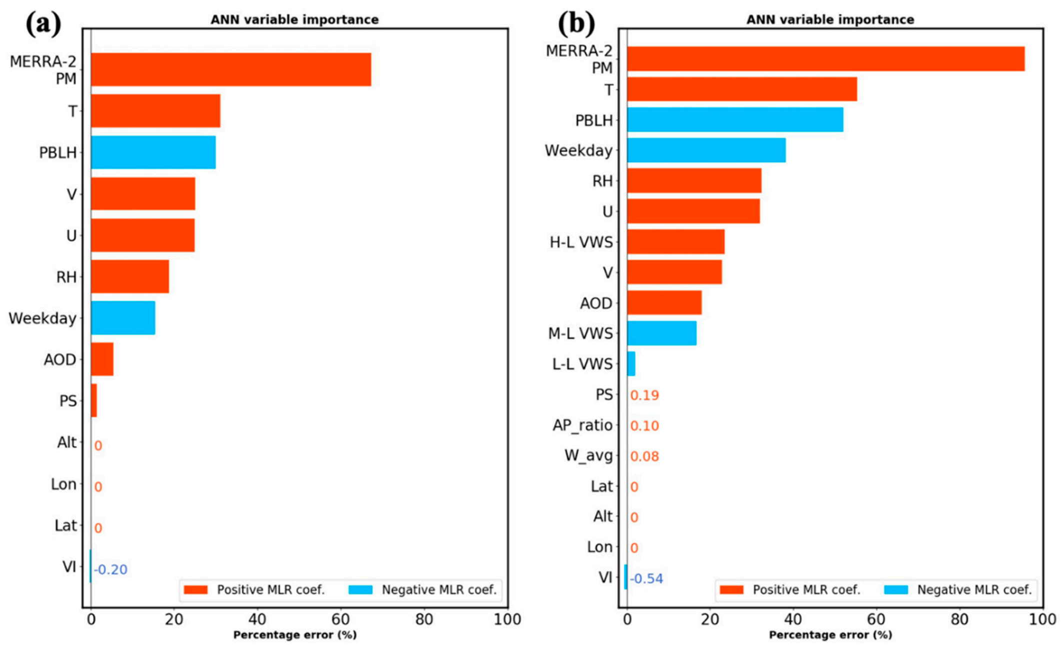

Figure 5. Variable importance of (**a**) set 1 and (**b**) set 2 models at PS 314 site. Values indicate the permutation importance estimated from the ANN models, and colors indicate the signs of regression coefficients in the MLR models.

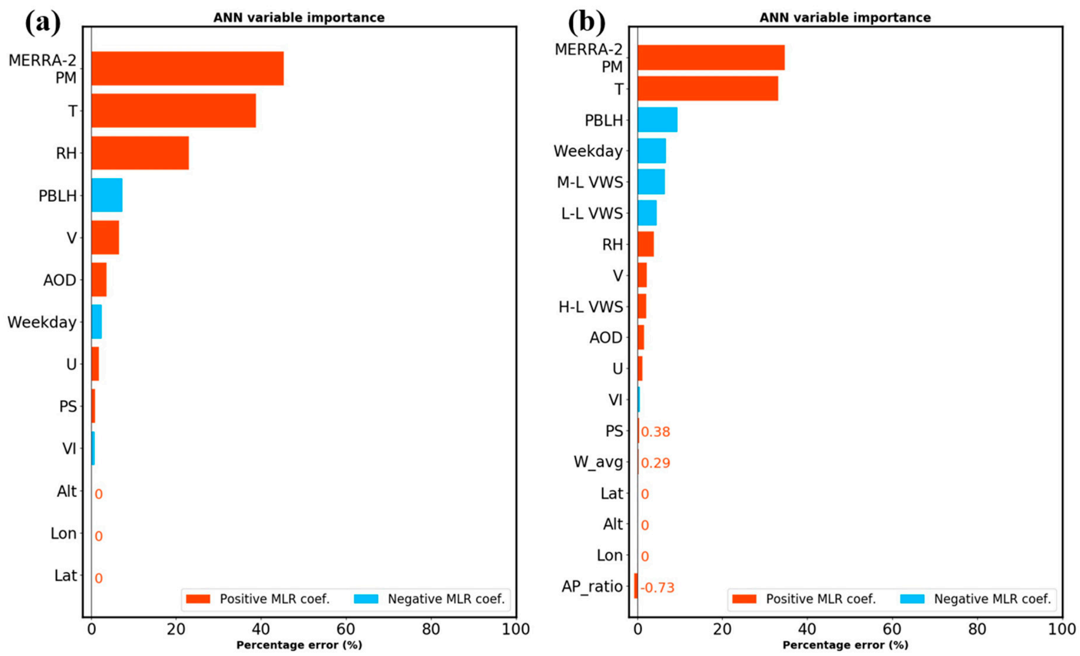

Figure 6. Variable importance of (**a**) set 1 and (**b**) set 2 models at Rochester site. Values indicate the permutation importance estimated from the ANN models, and colors indicate the signs of regression coefficients in the MLR models.

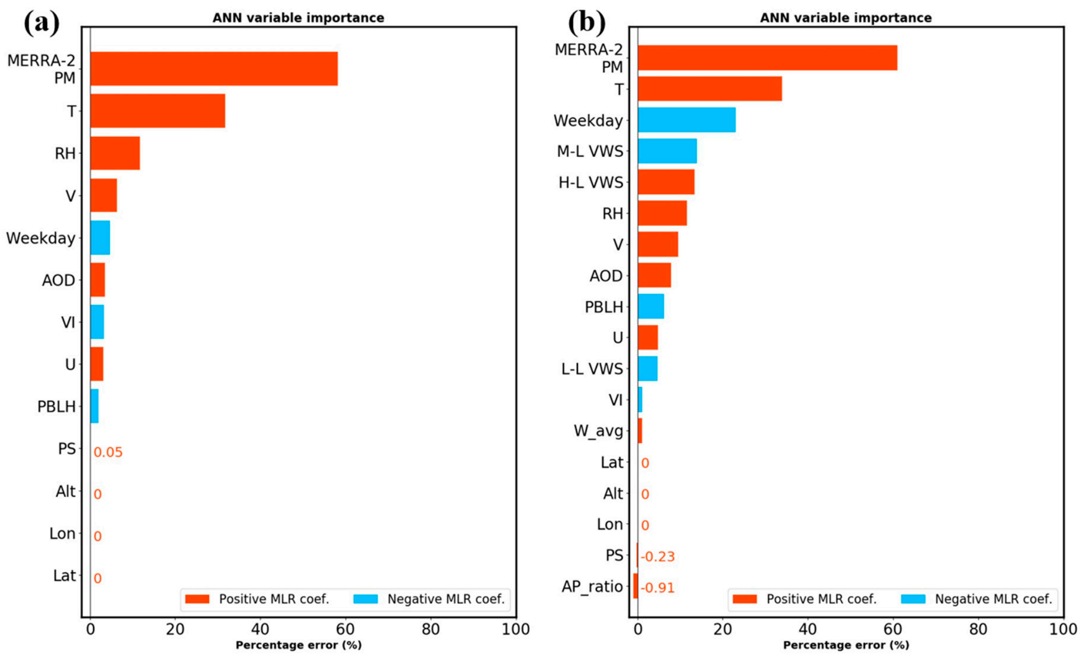

Figure 7. Variable importance of (**a**) set 1 and (**b**) set 2 models at Rockland County site. Values indicate the permutation importance estimated from the ANN models, and colors indicate the signs of regression coefficients in the MLR models.

For all sites, MERRA-2 $PM_{2.5}$ concentration and T showed the highest importance. The positive regression coefficients of MERRA-2 $PM_{2.5}$ concentration were related to the positive effects of aerosol removal processes and chemical reactions on $PM_{2.5}$ concentrations. The positive coefficients of T were consistent with previous studies [10,11], which showed warm condition was conductive for photochemical formation of secondary aerosols. Thus, the high importance of two predictors collectively indicated that aerosol removal processes and chemical reactions played a dominant role in $PM_{2.5}$ concentration during the daytime. Furthermore, the high importance of PBLH at the PS 314 and Rochester sites indicated the significant dispersion effect in the PBL. During the daytime, radiative heating on the surface led to a sharp increase in PBLH, resulting in decreasing the $PM_{2.5}$ concentration due to stronger dispersion. The weekday also showed significant importance at all sites. The negative coefficients between weekday and $PM_{2.5}$ concentrations may be associated with the anthropogenic emissions from industries and traffic during weekdays.

Additionally, surface wind fields, U and V, and VWSs showed moderate importance. The positive coefficients of U and V indicated the positive effects of westerly (positive U) and southerly (positive V) winds, respectively. This was consistent with previous studies, which showed that high $PM_{2.5}$ concentrations were associated with transported aerosols driven by westerly and southwesterly flows [41,42]. Moreover, the negative coefficients of M-VWS and L-VWS showed that weak air mass exchanges above the PBL and stable conditions in the PBL potentially led to high $PM_{2.5}$ concentrations, respectively. These results could be associated with the frontal systems caused by the westerly mid-latitude cyclones, which are common during summer in the Northeast US [42,66]. Additionally, the strong VWS at higher troposphere could be beneficial for the downward transport of long-range transported aerosols (e.g., smoke aerosols), resulting in the positive correlation between H-VWS and $PM_{2.5}$ concentrations.

At the PS 314 site (Figure 5), the VWS at three levels played a relatively weak role in $PM_{2.5}$ concentrations compared to the surface conditions. This was consistent with previous discussion, which mentioned that the influence of local ambient conditions on $PM_{2.5}$ concentrations were more significant than vertical mixing at NYC sites. In contrast, M-VWS and L-VWS at the Rochester site (Figure 6) showed comparable importance with surface winds and RH, indicating a similar, even stronger, influence of vertical mixing than surface ambient conditions. Similar phenomenon was also found at Rockland County site (Figure 7), with higher importance of M-VWS and H-VWS than surface winds, RH and PBLH. In addition, the low importance of PBLH and L-VWS at the Rockland County site may indicate a weak PBL-$PM_{2.5}$ correlation, since the $PM_{2.5}$ concentrations at rural sites are relatively low and not dominated by local emissions.

The positive correlation between AOD and $PM_{2.5}$ concentrations was expected. Since AOD is an indicator of total-column aerosol loadings, higher AOD represents higher $PM_{2.5}$ concentration particularly when most of the columnar aerosol loading is present in the PBL. According to previous studies, AOD showed significant impacts on $PM_{2.5}$ concentrations with high variable importance [32]. However, in this study, AOD had moderate importance and relatively weak influence compared to most of the other predictors. This could be due to the application of MERRA-2 $PM_{2.5}$ concentration. MERRA-2 is an aerosol reanalysis, which consists of both model simulations and observation assimilation. Therefore, MERRA-2 $PM_{2.5}$ concentration already includes AOD information from the assimilated AOD measurements. The AOD assimilation in MERRA-2 could contribute to the high importance of MERRA-2 $PM_{2.5}$ concentration. Since aerosol removal processes and chemical reactions were dominant during the daytime, MERRA-2 $PM_{2.5}$ concentration showed higher importance than AOD alone. The model results without MERRA-2 $PM_{2.5}$ concentration (not shown) also indicated the significant contributions of aerosol photochemical reactions with the highest importance of T.

Based on the definition, high AP_ratio reflects the presence of aloft aerosol layers and positive W_avg represents the atmospheric downward mixing. With strong downward mixings, aloft aerosols could descend to the surface, resulting in increasing the surface $PM_{2.5}$ concentrations. The positive coefficients at all sites indicated the positive effects of aloft aerosols on $PM_{2.5}$ concentrations. However, the low importance of the two predictors showed that the influence of aloft aerosols was relatively minor. It was probably because of the limited cases of the presence of high $PM_{2.5}$ concentrations and aloft aerosols, and the capability of AP_ratio in representing aloft aerosols. In addition, the low importance of PS could be due to its dependence on Alt. Atmospheric pressure provides information about circulation patterns, synoptic weather systems, and atmospheric stability conditions. Since PS decreases with height, such relation may dilute the connection between PS and meteorology, and the correlation between PS and $PM_{2.5}$ concentrations, especially for UNY and rural sites with a wide range of Alt values.

4. Conclusions

In NYS, episodic high $PM_{2.5}$ concentrations due to transported aerosols have been reported in summer. Driven by synoptic downward mixing and PBL entrainment, pollutants potentially transport from free troposphere into the PBL and affect $PM_{2.5}$ concentrations near the surface. In light of the contributions of transported aerosols to high $PM_{2.5}$ concentrations in NYS, understandings of the relationships between the vertical mixing of aloft aerosols and $PM_{2.5}$ concentrations are important. This study investigated the influences of various factors on the $PM_{2.5}$ concentrations in NYS by analyzing the testing results of multiple linear regression (MLR) and artificial neural network (ANN) models trained with two sets of predictors. Overall, the ANN models performed better than MLR models. Additionally, for the ANN models, the predictors of vertical mixing and aloft aerosols improved the MB, R^2 and RMSE by 0.34, 0.02 and 0.17 $\mu g\ m^{-3}$, respectively, on average. Although RMSEs around 3 $\mu g\ m^{-3}$ were relatively high for NYS, as the RMSEs were one-third and/or close to the annual $PM_{2.5}$ average concentrations of 2019, the improvements in model performance were non-negligible. The leave-one-out cross-validation results showed significant site-variations and were able to differentiate predictor-$PM_{2.5}$ correlations at sites with different air quality characteristics. The model improvement due to vertical mixing and aloft aerosols was more significant at rural sites, where the $PM_{2.5}$ concentrations are mainly affected by meteorology and transported aerosols. The changes in model performance at UNY sites varied among sites, probably due to the spatial variability within a wide region of UNY. In addition, a joint analysis of regression coefficients and variable importance provided insights into the contributions of selected predictors to $PM_{2.5}$ concentration. The aerosol removal process and chemical reactions showed the highest importance at three categories of sites (UNY, NYC and rural), and the contribution of vertical mixing was more significant at UNY and rural sites. However, the influence of aloft aerosols was limited in the current results. Identifying the cases of high $PM_{2.5}$ concentrations associated with transported aerosols prior to training may provide more significant results and better understanding of the influence of aloft aerosols.

Author Contributions: Conceptualization: W.-T.H., C.-H.L., S.A., R.K., C.-A.L.; Methodology: W.-T.H., C.-H.L., S.A., R.K., C.-A.L.; Software: W.-T.H., S.A.; Formal analysis: W.-T.H., C.-H.L., S.A., R.K.; Investigation: W.-T.H.; Writing—original draft preparation: W.-T.H.; Writing—review and editing: C.-H.L., R.K., C.-A.L.; Visualization: W.-T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the New York State Energy Research and Development Authority (NYSERDA) project (100417). The National Center for Atmospheric Research (NCAR) is supported by the National Science Foundation (NSF).

Acknowledgments: We thank our program manager, Ellen Burkhard (NYSERDA). We would also like to acknowledge high-performance computing support from Casper (https://www2.cisl.ucar.edu/resources/computational-systems/casper) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the NSF. The MERRA-2 data have been provided by the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center. We thank the NOAA/Earth System Research Laboratories (ESRL)/Global Systems Laboratory (GSL) and the Center for High Performance Computing (CHPC) at the University of Utah for providing archived HRRR data. We acknowledge the NOAA National Environmental Satellite, Data, and Information Service (NESDIS)/Center for Satellite Applications and Research (STAR) group for providing the VIIRS AOD data. The VIIRS VI data have been provided by the NASA's Land Processes Distributed Active Archive Center (LP DAAC). We also thank the US Environmental Protection Agency (EPA) for providing in situ $PM_{2.5}$ observations.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

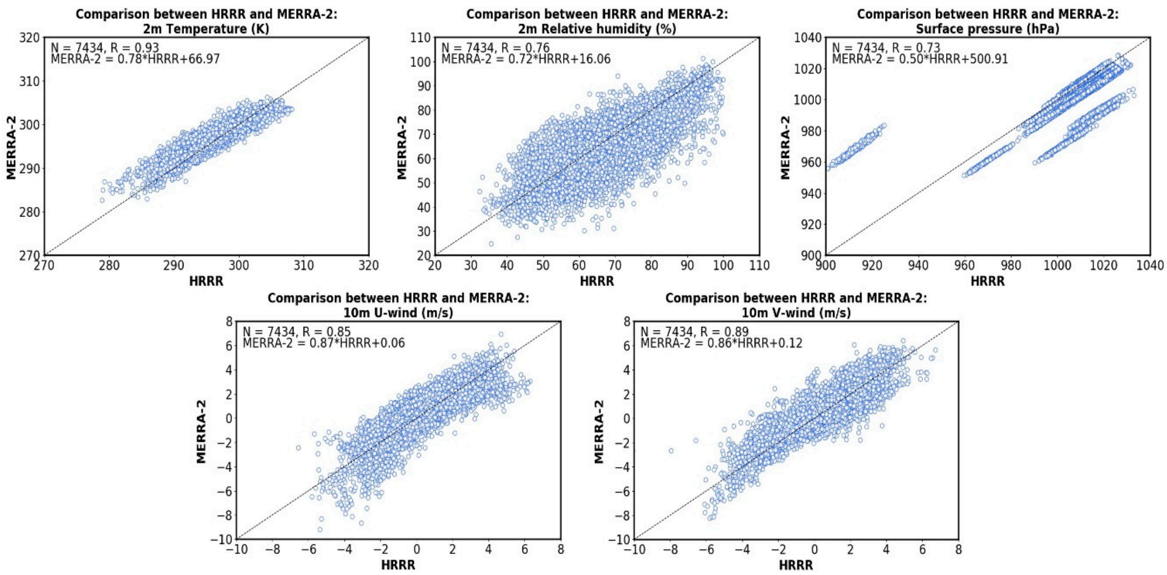

Figure A1. Comparison between HRRR and MERRA-2 meteorological fields.

Appendix B

Table A1. Correlation coefficients (R) between variables.

	U	V	RH	T	PBLH	PS	MERRA2_PM	Lat	Lon	VI
U										
V	−0.04									
RH	−0.22	0.30								
T	0.17	0.24	−0.14							
PBLH	0.34	−0.32	−0.66	0.21						
PS	−0.23	−0.10	−0.19	0.32	0.08					
MERRA2_PM	−0.03	0.26	0.19	0.39	−0.12	0.16				
Lat	0.17	0.04	0.14	−0.35	−0.05	−0.70	−0.27			
Lon	−0.15	−0.05	−0.02	0.13	0.00	0.25	0.15	−0.60		
VI	0.10	0.00	0.11	−0.20	0.04	−0.54	−0.18	0.62	−0.21	
Alt	0.15	0.07	0.16	−0.38	−0.06	−0.92	−0.23	0.70	−0.40	0.63
Weekday	−0.01	0.08	0.03	−0.03	0.00	−0.03	0.08	−0.01	0.01	0.00
H-VWS	0.14	−0.26	−0.18	−0.32	0.22	−0.02	−0.13	0.03	0.01	0.02
M-VWS	0.18	−0.22	−0.19	−0.20	0.20	−0.08	−0.24	0.07	−0.04	0.04
L-VWS	0.14	0.35	0.25	0.06	−0.28	0.02	0.16	−0.09	0.07	−0.11
W_avg	0.15	0.02	−0.04	0.14	0.05	0.29	0.09	−0.23	−0.07	−0.21
AP_ratio	−0.01	0.00	0.02	0.00	−0.01	0.00	0.04	0.00	−0.01	0.00
AOD	−0.08	0.24	0.27	0.31	−0.10	0.15	0.61	−0.30	0.20	−0.26
Obs_PM	0.06	0.38	0.19	0.49	−0.16	0.04	0.57	−0.05	−0.05	−0.12

	Alt	Weekday	H-VWS	M-VWS	L-VWS	W_avg	AP_ratio	AOD
U								
V								
RH								
T								
PBLH								
PS								
MERR2A_PM								
Lat								
Lon								
VI								
Alt								
Weekday	0.00							
H-VWS	0.01	−0.01						
M-VWS	0.06	−0.05	0.22					
L-VWS	−0.11	−0.04	−0.11	0.08				
W_avg	−0.27	0.01	−0.01	0.04	0.21			
AP_ratio	0.00	−0.01	0.01	0.02	0.01	0.00		
AOD	−0.23	0.09	−0.15	−0.20	0.14	0.07	0.00	
obs_PM	−0.09	0.02	−0.22	−0.26	0.13	0.06	0.03	0.40

H-VWS, M-VWS and L-VWS are the vertical wind shears of 700–500 hPa, 850–700 hPa and surface –850 hPa, respectively. W_avg is the average vertical velocity between surface and 500 hPa. AP_ratio is the ratio of AOD change rate to $PM_{2.5}$ change rate.

Appendix C

Table A2. LOOCV testing results of the MLR models at selected air quality sites.

Label	Name	ID Number	Bias (μg m^{-3})		R-Squared		RMSE (μg m^{-3})	
			Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
1	Albany	360010005	−1.93	−2.01	0.39	0.40	3.63	3.66
2	Buffalo	360290005	0.37	0.45	0.47	0.47	2.64	2.66
3	Tonawanda II	360291014	0.67	0.72	0.50	0.49	3.33	3.37
4	Rochester	360551007	−0.51	−0.53	0.45	0.45	2.83	2.83
5	Utica	360652001	1.19	1.12	0.41	0.40	3.22	3.21
6	Whiteface Mountain	360310003	2.36	3.47	0.54	0.52	3.22	4.13
7	Rockland County	360870005	−0.18	−0.19	0.55	0.56	2.83	2.82
8	Pinnacle State Park	361010003	−2.64	−3.06	0.56	0.55	3.29	3.65
9	Bronx	360050112	0.32	0.32	0.58	0.58	2.83	2.82
10	PS 314	360470052	1.11	1.08	0.56	0.57	2.68	2.65
11	PS 274	360470118	1.08	1.09	0.55	0.56	2.95	2.93
12	Esienhower Park	360590005	0.49	0.52	0.55	0.56	2.91	2.89
13	IS 143	360610115	−0.93	−0.92	0.58	0.59	2.94	2.92
14	Division St.	360610134	−0.40	−0.41	0.52	0.53	2.83	2.80
15	CCNY	360610135	−0.64	−0.64	0.50	0.51	2.88	2.86
16	Newburgh	360710002	0.69	0.65	0.45	0.46	2.65	2.63
17	Maspeth	360810120	0.84	0.84	0.54	0.55	2.88	2.87
18	Queens	360810124	−0.74	−0.76	0.59	0.60	2.74	2.72
19	FKILL	360850111	−1.06	−1.06	0.41	0.42	3.26	3.25
20	Holtsville	361030009	1.03	1.09	0.54	0.54	2.64	2.67
21	White Plain	361192004	−0.59	−0.52	0.54	0.54	2.91	2.88

Appendix D

Table A3. LOOCV testing results of the ANN models at selected air quality sites.

Label	Name	ID Number	Bias (μg m^{-3})		R-Squared		RMSE (μg m^{-3})	
			Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
1	Albany	360010005	−1.42	−1.52	0.57	0.55	2.97	3.07
2	Buffalo	360290005	−0.12	0.66	0.62	0.62	2.24	2.61
3	Tonawanda II	360291014	0.23	0.37	0.62	0.64	2.88	2.79
4	Rochester	360551007	−0.75	−0.26	0.63	0.62	2.40	2.33
5	Utica	360652001	0.51	0.12	0.50	0.50	2.79	2.75
6	Whiteface Mountain	360310003	−2.99	−1.16	0.57	0.57	3.89	2.38
7	Rockland County	360870005	−0.71	−0.10	0.66	0.71	2.63	2.26
8	Pinnacle State Park	361010003	−2.60	−1.80	0.60	0.63	3.22	2.55
9	Bronx	360050112	−0.60	0.72	0.73	0.75	2.35	2.24
10	PS 314	360470052	0.28	0.31	0.73	0.81	1.92	1.64
11	PS 274	360470118	1.25	1.28	0.73	0.79	2.47	2.27
12	Esienhower Park	360590005	0.21	0.34	0.69	0.71	2.47	2.44
13	IS 143	360610115	0.13	−0.16	0.69	0.74	2.38	2.20
14	Division St.	360610134	−0.87	−0.56	0.74	0.76	2.27	2.07
15	CCNY	360610135	−0.98	−1.01	0.70	0.75	2.40	2.25
16	Newburgh	360710002	−1.74	−0.70	0.46	0.49	3.10	2.59
17	Maspeth	360810120	0.14	0.14	0.76	0.79	2.05	1.89
18	Queens	360810124	−0.83	−1.36	0.78	0.78	2.09	2.38
19	FKILL	360850111	−1.15	−2.00	0.54	0.57	2.97	3.32
20	Holtsville	361030009	−0.39	−0.21	0.61	0.62	2.25	2.21
21	White Plain	361192004	−0.88	0.73	0.66	0.68	2.71	2.46

Appendix E

Table A4. Regression coefficients from the MLR models at PS 314, Rochester and Rockland County sites.

Site	U	V	RH	T	PBLH (10^{-3})	PS (10^{-4})	MERRA2_PM	Lat	Lon	VI	Alt
MLR-1											
PS 314	0.088	0.246	0.012	0.437	−1.587	3.611	0.277	1.025	0.068	−5.370	0.007
Rochester	0.101	0.248	0.009	0.424	−1.556	3.568	0.289	0.982	0.075	−4.764	0.007
Rockland County	0.114	0.236	0.012	0.431	−1.586	3.444	0.277	1.018	0.055	−4.851	0.006
MLR-2											
PS 314	0.109	0.264	0.012	0.437	−1.678	3.269	0.271	1.005	0.063	−5.298	0.006
Rochester	0.122	0.271	0.010	0.427	−1.689	3.239	0.283	0.959	0.071	−4.715	0.006
Rockland County	0.137	0.257	0.013	0.430	−1.690	3.040	0.272	0.997	0.052	−4.790	0.006

Site	Weekday	AOD	H-VWS	M-VWS	L-VWS	W_avg	AP_ratio
MLR-1							
PS 314	−0.008	1.342					
Rochester	−0.008	1.062					
Rockland County	−0.004	1.111					
MLR-2							
PS 314	−0.015	1.357	91.666	−118.698	−83.530	0.625	0.009
Rochester	−0.015	1.080	105.872	−105.617	−102.905	0.452	0.007
Rockland County	−0.012	1.125	88.544	−117.948	−100.918	0.945	0.010

References

1. Pope, I.C.; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J. Am. Med. Assoc.* **2002**, *287*, 1132–1141. [CrossRef] [PubMed]
2. Pope, C.A.; Burnett, R.T., III; Thurston, G.D.; Thun, M.J.; Calle, E.E.; Krewski, D.; Godleski, J.J. Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution: Epidemiological Evidence of General Pathophysiological Pathways of Disease. *Circulation* **2004**, *1*, 71–77. [CrossRef] [PubMed]
3. Apte, J.S.; Marshall, J.D.; Cohen, A.J.; Brauer, M. Addressing Global Mortality from Ambient PM2.5. *Environ. Sci. Technol.* **2015**, *49*, 8057–8066. [CrossRef] [PubMed]
4. Franklin, M.; Zeka, A.; Schwartz, J. Association between PM2.5 and all-cause and specific-cause mortality in 27 US communities. *J. Expo. Sci. Environ. Epidemiol.* **2007**, *17*, 279–287. [CrossRef]
5. Behera, S.N.; Sharma, M. Reconstructing Primary and Secondary Components of PM2.5 Composition for an Urban Atmosphere. *Aerosol Sci. Technol.* **2010**, *44*, 983–992. [CrossRef]
6. Edney, E.O.; Kleindienst, T.E.; Jaoui, M.; Lewandowski, M.; Offenberg, J.H.; Wang, W.; Claeys, M. Formation of 2-methyltetrols and 2-methylglyceric acid in secondary organic aerosol from laboratory irradiated isoprene/NO_X/SO_2/air mixtures and their detection in ambient PM2.5 samples collected in the eastern United States. *Atmos. Environ.* **2005**, *39*, 5281–5289. [CrossRef]
7. Lonati, G.; Ozgen, S.; Giugliano, M. Primary and secondary carbonaceous species in PM2.5 samples in Milan (Italy). *Atmos. Environ.* **2007**, *41*, 4599–4610. [CrossRef]
8. Wang, Y.; Zhuang, G.; Tang, A.; Yuan, H.; Sun, Y.; Chen, S.; Zheng, A. The ion chemistry and the source of PM2.5 aerosol in Beijing. *Atmos. Environ.* **2005**, *39*, 3771–3784. [CrossRef]
9. Yu, S.; Mathur, R.; Schere, K.; Kang, D.; Pleim, J.; Young, J.; Tong, D.; Pouliot, G.; McKeen, S.A.; Rao, S.T. Evaluation of real-time PM2.5 forecasts and process analysis for PM2.5 formation over the eastern United States using the Eta-CMAQ forecast model during the 2004 ICARTT study. *J. Geophys. Res.* **2008**, *113*, D06204. [CrossRef]
10. Dawson, J.P.; Adams, P.J.; Pandis, S.N. Sensitivity of PM2.5 to climate in the Eastern US: A modeling case study. *Atmos. Chem. Phys.* **2007**, *7*, 4295–4309. [CrossRef]
11. Tai, A.P.K.; Mickley, L.J.; Jacob, D.J. Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *Atmos. Environ.* **2010**, *44*, 3976–3984. [CrossRef]
12. Tran, H.N.Q.; Mölders, N. Investigations on meteorological conditions for elevated PM2.5 in Fairbanks, Alaska. *Atmos. Res.* **2011**, *99*, 39–49. [CrossRef]

13. Wang, J.; Ogawa, S. Effects of Meteorological Conditions on PM2.5 Concentrations in Nagasaki, Japan. *Int. J. Environ. Res. Public Health* **2015**, *12*, 9089–9101. [CrossRef] [PubMed]
14. Zhang, Z.; Zhang, X.; Gong, D.; Quan, W.; Zhao, X.; Ma, Z.; Kim, S.-J. Evolution of surface O_3 and PM2.5 concentrations and their relationships with meteorological conditions over the last decade in Beijing. *Atmos. Environ.* **2015**, *108*, 67–75. [CrossRef]
15. Chen, Y.; An, J.; Wang, X.; Sun, Y.; Wang, Z.; Duan, J. Observation of wind shear during evening transition and an estimation of submicron aerosol concentrations in Beijing using a Doppler wind lidar. *J. Meteor. Res.* **2017**, *31*, 350–362. [CrossRef]
16. Li, Z.; Guo, J.; Ding, A.; Liao, H.; Liu, J.; Sun, Y.; Wang, T.; Xue, H.; Zhang, H.; Zhu, B. Aerosol and boundary-layer interactions and impact on air quality. *Nat. Sci. Rev.* **2017**, *4*, 810–833. [CrossRef]
17. Yang, Y.; Yim, S.H.L.; Haywood, J.; Osborne, M.; Chan, J.C.S.; Zeng, Z.; Cheng, J.C.H. Characteristics of heavy particulate matter pollution events over Hong Kong and their relationships with vertical wind profiles using high-time-resolution Doppler lidar measurements. *J. Geophys. Res. Atmos.* **2019**, *124*, 9609–9623. [CrossRef]
18. Zhang, J.; Rao, S.T. The Role of Vertical Mixing in the Temporal Evolution of Ground-Level Ozone Concentrations. *J. Appl. Meteor.* **1999**, *38*, 1674–1691. [CrossRef]
19. Zhang, Y.; Guo, J.; Yang, Y.; Wang, Y.; Yim, S.H.L. Vertical Wind Shear Modulates Particulate Matter Pollutions: A Perspective from Radar Wind Profiler Observations in Beijing, China. *Remote Sens.* **2020**, *12*, 546. [CrossRef]
20. Hu, J.; Chen, J.; Ying, Q.; Zhang, H. One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system. *Atmos. Chem. Phys.* **2016**, *16*, 10333–10350. [CrossRef]
21. Saide, P.E.; Carmichael, G.R.; Spak, S.N.; Gallardo, L.; Osses, A.E.; Mena-Carrasco, M.A.; Pagowski, M. Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. *Atmos. Environ.* **2011**, *45*, 2769–2780. [CrossRef]
22. Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dėdelė, A.; Dons, E.; De Nazelle, A.; et al. Development of Land Use Regression Models for PM2.5, PM2.5 Absorbance, PM10 and PMcoarse in 20 European Study Areas; Results of the ESCAPE Project. *Environ. Sci. Technol.* **2012**, *46*, 11195–11205. [CrossRef] [PubMed]
23. Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008**, *42*, 7561–7578. [CrossRef]
24. Chu, Y.; Liu, Y.; Li, X.; Liu, Z.; Lu, H.; Lu, Y.; Mao, Z.; Chen, X.; Li, N.; Ren, M.; et al. A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. *Atmosphere* **2016**, *7*, 129. [CrossRef]
25. Liu, Y.; Sarnat, J.A.; Kilaru, V.; Jacob, D.J.; Koutrakis, P. Estimating Ground-Level PM2.5 in the Eastern United States Using Satellite Remote Sensing. *Environ. Sci. Technol.* **2005**, *39*, 3269–3278. [CrossRef] [PubMed]
26. Van Donkelaar, A.; Martin, R.V.; Spurr, R.J.D.; Burnett, R.T. High-Resolution Satellite-Derived PM2.5 from Optimal Estimation and Geographically Weighted Regression over North America. *Environ. Sci. Technol.* **2015**, *49*, 10482–10491. [CrossRef]
27. Gupta, P.; Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *J. Geophys. Res.* **2009**, *114*, D14205. [CrossRef]
28. Gupta, P.; Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. *J. Geophys. Res.* **2009**, *114*, D20205. [CrossRef]
29. Reid, C.E.; Jerrett, M.; Petersen, M.L.; Pfister, G.G.; Morefield, P.E.; Tager, I.B.; Raffuse, S.M.; Balmes, J.R. Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. *Environ. Sci. Technol.* **2015**, *49*, 3887–3896. [CrossRef]
30. Engel-Cox, J.A.; Holloman, C.H.; Coutant, B.W.; Hoff, R.M. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmos. Environ.* **2004**, *38*, 2495–2509. [CrossRef]
31. Weber, S.A.; Engel-Cox, J.A.; Hoff, R.M.; Prados, A.I.; Zhang, H. An improved method for estimating surface fine particle concentrations using seasonally adjusted satellite aerosol optical depth. *J. Air Waste Manag. Assoc.* **2010**, *60*, 574–585. [CrossRef] [PubMed]
32. Xu, Y.; Ho, H.C.; Wong, M.S.; Deng, C.; Shi, Y.; Chan, T.-C.; Knudby, A. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environ. Pollut.* **2018**, *242*, 1417–1426. [CrossRef] [PubMed]

33. Xue, T.; Zheng, Y.; Tong, D.; Zheng, B.; Li, X.; Zhu, T.; Zhang, Q. Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. *Environ. Int.* **2019**, *123*, 345–357. [CrossRef] [PubMed]
34. Yao, J.; Brauer, M.; Raffuse, S.; Henderson, S.B. Machine Learning Approach to Estimate Hourly Exposure to Fine Particulate Matter for Urban, Rural, and Remote Populations during Wildfire Seasons. *Environ. Sci. Technol.* **2018**, *52*, 13239–13249. [CrossRef] [PubMed]
35. Emami, F.; Masiol, M.; Hopke, P.K. Air pollution at Rochester, NY: Long-term trends and multivariate analysis of upwind SO_2 source impacts. *Sci. Total Environ.* **2018**, *612*, 1506–1515. [CrossRef] [PubMed]
36. Rattigan, O.V.; Felton, H.D.; Bae, M.-S.; Schwab, J.J.; Demerjian, K.L. Multi-year hourly PM2.5 carbon measurements in New York: Diurnal, day of week and seasonal patterns. *Atmos. Environ.* **2010**, *44*, 2043–2053. [CrossRef]
37. Rattigan, O.V.; Civerolo, K.L.; Felton, H.D.; Schwab, J.J.; Demerjian, K.L. Long Term Trends in New York: PM2.5 Mass and Particle Components. *Aerosol Air Qual. Res.* **2015**, *16*, 191–1203. [CrossRef]
38. Squizzato, S.; Masiol, M.; Rich, D.Q.; Hopke, P.K. PM2.5 and gaseous pollutants in New York state during 2005–2016: Spatial variability, temporal trends, and economic influences. *Atmos. Environ.* **2018**, *183*, 209–224. [CrossRef]
39. Bari, A.; Dutkiewicz, V.A.; Judd, C.D.; Wilson, L.R.; Luttinger, D.; Husain, L. Regional sources of particulate sulfate, SO_2, PM2.5, HCL, HNO_3, HONO and NH_3 in New York, NY. *Atmos. Environ.* **2003**, *37*, 2837–2844. [CrossRef]
40. Qin, Y.; Kim, E.; Hopke, P.K. The concentrations and sources of PM2.5 in metropolitan New York City. *Atmos. Environ.* **2006**, *40*, 312–332. [CrossRef]
41. DeGaetano, A.T.; Doherty, O.M. Temporal, spatial and meteorological variations in hourly PM2.5 concentration extremes in New York City. *Atmos. Environ.* **2004**, *38*, 1547–1558. [CrossRef]
42. Dutkiewicz, V.A.; Das, M.; Husain, L. The relationship between regional SO_2 emissions and downwind aerosol sulfate concentrations in the Northeastern US. *Atmos. Environ.* **2000**, *34*, 1821–1832. [CrossRef]
43. Dutkiewicz, V.A.; Qureshi, S.; Khan, A.R.; Ferraro, V.; Schwab, J.; Demerjian, K.; Husain, L. Sources of fine particulate sulfate in New York. *Atmos. Environ.* **2004**, *38*, 3179–3189. [CrossRef]
44. Dutkiewicz, V.A.; Husain, L.; Roychowdhury, U.K.; Demerjian, K.L. Impact of Canadian wildfire smoke on air quality at two rural sites in NY State. *Atmos. Environ.* **2011**, *45*, 2028–2033. [CrossRef]
45. Hung, W.-T.; Lu, C.-H.S.; Shrestha, B.; Lin, H.-C.; Lin, C.-A.; Grohan, D.; Hong, J.; Ahmadov, R.; James, E.; Joseph, E. The impacts of transported wildfire smoke aerosols on surface air quality in New York State: A case study in summer 2018. *Atmos. Environ.* **2020**, *227*, 117415. [CrossRef]
46. Roger, H.M.; Ditto, J.C.; Gentner, D.R. Evidence for impacts on surface-level air quality in the northeastern US from long-distance transport of smoke from North American fires during the Long Island Sound Tropospheric Ozone Study (LISTOS) 2018. *Atmos. Chem. Phys.* **2020**, *20*, 671–682. [CrossRef]
47. Wu, Y.; Arapi, A.; Huang, J.; Gross, B.; Moshary, F. Intra-continental wildfire smoke transport and impact on local air quality observed by ground-based and satellite remote sensing in New York City. *Atmos. Environ.* **2018**, *187*, 266–281. [CrossRef]
48. Zu, K.; Tao, G.; Long, K.; Goodman, J.; Valberg, P. Long-range fine particulate matter from the 2002 Quebec forest fires and daily mortality in Greater Boston and New York City. *Air Qual. Atmos. Health* **2016**, *9*, 213–221. [CrossRef]
49. Alexander, C.R.; Weygandt, S.S.; Smirnova, T.G.; Benjamin, S.; Hofmann, P.; James, E.P.; Koch, D.A. High Resolution Rapid Refresh (HRRR): Recent enhancements and evaluation during the 2010 convective season. In Proceedings of the 25th Conference on Severe Local Storms, Denver, CO, USA, 12 October 2010.
50. Cao, C.; Deluccia, F.; Xiong, X.; Wolfe, R.; Weng, F. Early on-orbit performance of the VIIRS onboard the S-NPP satellite. *IEEE Trans. Geosci. Remote Sens.* **2013**, *99*. [CrossRef]
51. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Climate* **2017**, *30*, 5419–5454. [CrossRef]
52. Colarco, P.; da Silva, A.; Chin, M.; Diehl, T. Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth. *J. Geophys. Res.* **2010**, *115*, D14207. [CrossRef]

53. 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.; et al. The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies. *J. Climate* **2017**, *30*, 6851–6872. [CrossRef] [PubMed]
54. Randles, C.A.; da Silva, A.M.; Buchard, V.; Colarco, P.R.; Darmenov, A.; Govindaraju, R.; Smirnov, A.; Holben, B.; Ferrare, R.; Hair, J.; et al. The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System description and data assimilation evaluation. *J. Climate* **2017**, *30*, 6823–6850. [CrossRef]
55. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
56. Obata, K.; Miura, T.; Yoshioka, H.; Huete, A.R.; Vargas, M. Spectral Cross-Calibration of VIIRS Enhanced Vegetation Index with MODIS: A Case Study Using Year-Long Global Data. *Remote Sens.* **2016**, *8*, 34. [CrossRef]
57. Géron, A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media: Sebastopol, CA, USA, 2017; ISBN 978-1491962299.
58. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
59. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
60. Fletcher, D.; Goss, E. Forecasting with neural networks: An application using bankruptcy data. *Inf. Manag.* **1993**, *24*, 159–167. [CrossRef]
61. Chollet, F. *Deep Learning with Python*; Manning Publications Co.: Greenwich, CT, USA, 2017; ISBN 9781617294433.
62. Sammut, C.; Webb, G.I. *Encyclopedia of Machine Learning*; Springer Publishing Company: New York, NY, USA, 2011; ISBN 9780387301648.
63. Watson, G.; Telesca, D.; Reid, C.; Pfister, G.; Jerrett, M. Machine learning models accurately model ozone exposure during wildfire events. *Environ. Pollut.* **2019**, *254*. [CrossRef]
64. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
65. McGovern, A.; Lagerquist, R.; Gagne, D.J.; Jergensen, G.E.; Elmore, K.L.; Homeyer, C.R.; Smith, T. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bull. Amer. Meteor. Soc.* **2019**, *100*, 2175–2199. [CrossRef]
66. Pal, S.; Davis, K.J.; Lauvaux, T.; Browell, E.V.; Gaudet, B.J.; Stauffer, D.R.; Obland, M.D.; Choi, Y.; DiGangi, J.P.; Feng, S.; et al. Observations of greenhouse gas changes across summer frontal boundaries in the eastern United States. *J. Geophys. Res. Atmos.* **2020**, *125*, e2019JD030526. [CrossRef]