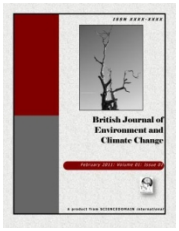

British Journal of Environment & Climate Change
3(4): 547-565, 2013

SCIENCEDOMAIN *international*
www.sciencedomain.org

Assessing Surface PM2.5 Estimates Using Data Fusion of Active and Passive Remote Sensing Methods

Lina Cordero[1*], Nabin Malakar[1], Yonghua Wu[1], Barry Gross[1] and Fred Moshary[1]

[1]Optical Remote Sensing Lab, City College of New York, New York, NY, USA 10031, USA.

Authors' contributions

This work was carried out in collaboration between all authors. Finally all authors read and approved the final manuscript.

Original Research Article

Received 2nd November 2013
Accepted 2nd November 2013
Published 18th November 2013

ABSTRACT

In this paper, we focus on estimations of fine particulate matter by combining MODIS satellite Aerosol Optical Depth (AOD) with Weather Research Forecast (WRF) PBL information using a neural network approach and assess its performance. As part of our analysis, we first explore the baseline effectiveness of AOD and PBL as relevant factors in estimating PM2.5 in passive radiometer and active lidar data at CCNY and demonstrate that the PBL height is the most critical additional parameter for accurate PM2.5. Furthermore, active measurements from both ground and satellite based lidar are used to show that summer WRF model PBL heights are most accurate. We then expand our analysis to a regional domain where daily estimations are obtained and compared with operational GEOS-CHEM PM2.5 product. Using our approach, we also create regional daily PM2.5 maps and compare against GEOS-CHEM outputs. Finally, we also consider additional improvements, where multiple satellite observations are used as regressors to predict PM2.5. These results illustrate the significant improvement we obtain within this framework in comparison to a "one size fits all continental scale approach".

Keywords: PM2.5; AOD; PBL; LIDAR; CMAQ; WRF; GEOS-CHEM; MODIS; air quality.

**Corresponding author: Email: lina.cordero@gmail.com;*

ACRONYMS

***AERONET**: AErosol RObotic NETwork: **AOD**: Aerosol Optical Depth; **CALIOP**: The Cloud-Aerosol Lidar with Orthogonal Polarization; **CALIPSO**: Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations; **CCNY**: The City College of New York; **CMAQ**: Community Multiscale Air Quality; **DEM**: Digital elevation in meters; **EPA**: United States Environment Protection Agency; **IDEA**: Infusing satellite Data into Environmental air quality Applications; **IDW**: inverse distance weighting; **LIDAR**: Light detection and ranging; **MODIS**: Moderate Resolution Imaging Spectroradiometer; **NASA**: National Aeronautics and Space Administration; **NN**: Neural network; **NYSDEC**: New York State Department of Environmental Conservation; **NYSERDA**: New York State Energy Research and Development Authority; **PBL**: Planetary boundary layer; **PM**: Particulate matter; **PM2.5**: Fine particulate matter with particle diameters less than 2.5 microns; **RH**: relative humidity; **RSIG**: Remote Sensing Information Gateway; **TEOM**: Tapered Element Oscillating Microbalance; **WRF**: Weather Research Forecast*

1. INTRODUCTION

The quantification of fine scale particulate matter is a major concern to the health community because it can be easily inhaled deep into the lungs, resulting in oxidative inflammation in vital organs. In particular, fine particulate matter with particle diameters less than 2.5 microns (PM2.5) has been linked to respiratory and pulmonary difficulties and for this reason, strong concentration guidelines have been developed by the U.S. Environmental Protection Agency (EPA) to limit exposure [1-3].

To assess compliance, the EPA (as well as relevant state agencies) generally rely on air quality measurements at the surface through specialized and expensive ground-based monitors which unfortunately limits the spatial extent of a quality controlled air quality network such as AIR Now. In order to overcome this limitation and to provide useful spatial distributions of surface level PM2.5, satellite remote sensing of aerosol properties has become a major tool.

In particular, significant efforts have been made to connect Aerosol Optical Depth (AOD) which is a measure of the path integrated aerosol extinction (i.e. opacity measure) to estimate ground-level PM2.5. For example, Zhang, et al. [4] performed a geographical comparison over the 10 EPA regions across the United States using AOD. They used seasonal regression relations for each region to estimate the PM2.5 from AOD retrievals. However, an “exact” relationship cannot be derived due to a wide range of factors, such as the large variability of aerosols, the effects of meteorology and the vertical structure of aerosols which is often (but not always) constrained by the Planetary Boundary Layer (PBL) height. In particular, if aerosols are well mixed and trapped within the PBL, it is very reasonable that the surface PM2.5 should be sensitive to the PBL height. In fact, we would expect a generally inverse relationship between PBL height and PM2.5 for a fixed AOD. Such conditions can be expected to occur in cases where convective heating dominates mechanical shear (eg. urban heat island). However, such behavior is not always met under real conditions. In a study conducted in Taiwan, Tsai et al. [5] studied seasonal variation in the relationship between PM2.5 and AOD. Unlike our analysis where correlations were strongest in the summer, they found higher correlations in the fall which they attribute to a relatively stable and well mixed boundary layer, whereas lower correlations found in the summer were attributed to the strong convection associated with unstable weather systems

over this period dominated by thunderstorms and typhoons [5]. These patterns are clearly not the same as those of the US North East where urban heating leads to much more stable conditions in the summer. Another study conducted in France by Boyouk et al. [6] found improved PM2.5–AOD correlations when the mixed boundary layer was considered. Moreover, the incorporation of meteorological variables, such as temperature (TEMP) and relative humidity (RH), into developed models also improved the PM2.5 prediction [7], and according to Gupta et al. [8], the best correlation between fine particulate matter (PM2.5) and AOD is seen when the PBL height is small and when the relative humidity is less than 50%. However, Schaap et al. [9] pointed out that the PM2.5–AOD correlation increased when comparisons were made at mid-day, suggesting that aerosols were much better mixed in the boundary layer during that time. It is also expected that relative humidity should also have an impact on the AOD (and this AOD to PM2.5 factors) via an increase in the size of the particles and a change in the refractive index (Hänel) [10].

In another approach, the use of highly sophisticated models as opposed to statistical tools has been attempted to account for factors such as relative humidity modification, aerosol speciation in homogeneity and complex PBL height dynamics [11-13]. As a prototypical example, the global model (GEOS-CHEM) is being used to estimate, on a daily basis, the spatial relationship between PM2.5 forecast and column path AOD which can then be used in conjunction with satellite AOD retrievals.

The GEOS-CHEM PM2.5/AOD approach has been extensively developed as part of now operational IDEA (Infusing satellite Data into Environmental air quality Applications) product and has matured to the level that real time spatial maps of PM2.5 are operationally available [14]. In summary, the IDEA algorithm approach uses fused MODIS Terra-Aqua satellite retrievals to calculate an average AOD product which is combined with low resolution GEOS-CHEM PM2.5/AOD ratios. Unfortunately, the low spatial resolution (0.5 deg) reduced the effectiveness of this approach in dealing with urban-suburban domains.

Fig. 1 illustrates the GEOS-CHEM PM2.5/AOD and its comparison to the CMAQ PM2.5/AOD for the summer period. While many of the spatial structures in the image are similar, it is clear that the magnitude of the ratios is very different and the GEOS-CHEM has a broader variability in the PM2.5/AOD ratio which can manifest itself in higher bias when standard bias correction is applied.

It is clear that the aerosol optical property estimation between the models is not consistent. Part of the poor model agreement likely stems from the fact that high resolution models, such as the WRF/CMAQ (Weather Research Forecast/Community Multi scale Air Quality) model strongly couples meteorological factors, PBL height and surface boundary conditions including emission inventories to estimate particulate concentrations and vertical distributions [15,16]. In addition, the humidification of the aerosols is very difficult to treat rigorously. To see more clearly, the differences between the use of CMAQ or GEOS-CHEM, we plot in Fig. 2, the statistical histograms for summer conditions. Here, we see that the GEOS-CHEM results are very biased in comparison to the ratio between the AIR Now sites and the satellite AOD. Much better agreement is seen when compared against the CMAQ model.

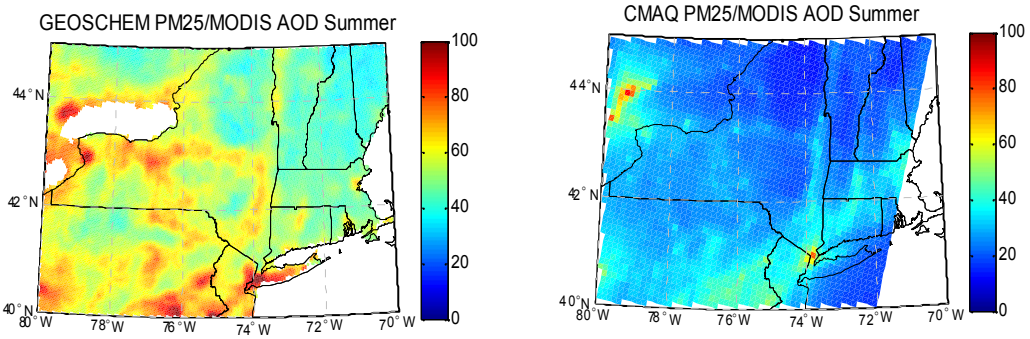

Fig. 1. Spatial comparison of PM2.5/AOD ratios (ug/m3 AOD) for different models (a) GEOS-CHEM (b) CMAQ New York State

Unfortunately, using high resolution AIR Quality models is very time and computer resource consuming and are not generally available in real time. On the other hand, using the WRF meteorological retrievals on their own is less time consuming and operationally simpler since the WRF modeling is generally much more straight-forward. Therefore, combining the satellite AOD with PBL height data seems to be a sensible compromise between the statistical and the fully coupled model approach. In addition, it should be pointed out that the model estimates of PM2.5 to AOD factors are strongly dependent on the quality of aerosol microphysics and adjustments to RH that are used. Unfortunately, it is generally agreed that the nature of the aerosols and the microphysical size distributions are very difficult to model making complete reliance on these models for PM-AOD factors less than ideal.

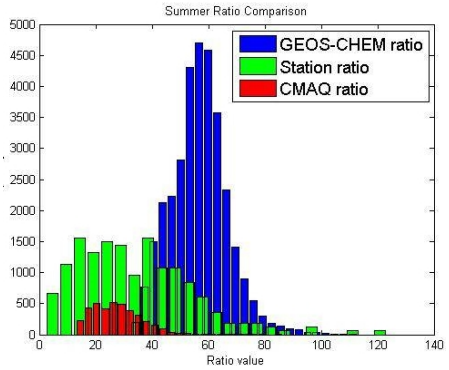

Fig. 2. Statistical Comparison of PM2.5/AOD ratios for different models in comparison to field station

However, to use model based PBL heights, we need to assess the potential of the models to qualitatively and quantitatively measure reasonable PBL heights. It is this particular demand that forces us to go beyond passive satellite remote sensing alone and to use active lidar based retrieval methods including ground based lidar at the City College of New York (CCNY) as well as spaced based active lidar (CALIOP) on board the CALIPSO platform. It is the combination of these tools that allow us to better understand when and why PBL information may be of use in improving PM2.5 estimates from AOD. The focus of this paper is to illustrate the application of data fusion from active and passive remote sensing from both ground based platforms as well as satellites instruments to synergistically provide insights into the estimation of PM2.5 in local as well as regional level. Furthermore, this paper specifically focuses on the New York state area, where data from several ground stations was available. In section 2, the general methodology involved in assessing PM2.5 against ground truth is discussed. In section 3, results pertaining to local relationships between ground-based AOD and lidar derived PBL heights with surface PM2.5 using combined active and passive remote sensing are given. In section 4, the use of WRF based PBL heights and satellite AOD's are explored within a neural network (NN) approach and additional factors such as seasonality and spatial location are considered. In section 5,

further enhancements using additional satellite measurements are explored. Conclusions and future directions will be given in section 6.

2. DATA AND METHODOLOGY

A total of 41 stations are used for urban/non-urban seasonal comparisons between ground PM2.5 and satellite/model PM2.5. Fig. 3 illustrates the locations of the stations as well as the urban classification, which is primarily based on site location; mostly, stations located in the New York City metro area are depicted as urban. Appendix 1 lists the stations with their geographical coordinates as well as the urban classification (urban = 1, non-urban = 0) and digital elevation (DEM) in meters. We focus on the NY State in this paper since the research is supported by the New York State Energy Research and Development Authority (NYSERDA) but the general approach should be expected to be reasonably compatible to wider domains.

In order to assess the PM2.5-AOD relationship on a regional scale, we have collected satellite AOD collection 5 data from the Moderate Resolution Imaging Spectroradiometer (MODIS) on board of TERRA and AQUA for the period corresponding from January 1, 2006 to December 31, 2007. Also, hourly PM2.5 data from the New York State Department of Environmental Conservation (NYSDEC) have been collected along with WRF/CMAQ planetary boundary layer height from the Remote Sensing Information Gateway (RSIG, URL:http://ofmpub.epa.gov/rsig/rsigserver?rsig2D.html) of the United States Environmental Protection Agency (EPA). Datasets are converted to daily averages following EPA regulations for further comparisons with GEOS-CHEM products. In order to assess the model PBL, data from the satellite CALIPSO were also used.

On the local scale, AOD as well as fine mode contributions are acquired from an AERONET (AErosol RObotic NETwork) Cimel sun/sky radiometer (CE-318) [17-19]. The PBL height is directly obtained using a Lidar (Light Detection and Ranging) instrument collocated at the City College of New York [20,21] and the PM2.5 is obtained on an hourly period from a TEOM (Tapered Element Oscillating Microbalance) instrument. Finally, meteorological information (relative humidity, wind speed, etc.) are collected and used to perform additional factor comparisons between ground measurements and remote sensing data.

In developing useful PM2.5 estimators on a regional scale, we will use neural networks to study the relationship between PM2.5 and AOD, which is known to be complex and nonlinear. Neural networks have proven to perform well in different areas of study, including atmospheric sciences where many complex relationships cannot be sufficiently understood by using statistical approaches [22-24]. Our focal point is the application of neural network method for improved PM2.5 estimation over long time periods, while at the same time investigating other dependences such as additional factors or seasonal changes. When training the neural network, we split the data in the interleaved mode to ensure that all sectors of data are chosen. We used Bayesian regulation backpropagation method [25], which is done by the 'trainbr' function provided by the neural network

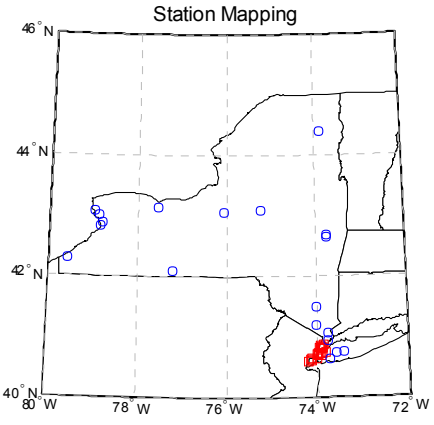

Fig. 3. New York State - Station Mapping: red squares are urban stations and blue circles are non-urban stations.

toolbox in Matlab (http://www.mathworks.com/products/neural-network/index.html). The training function updates the weight and bias values according to the Levenberg-Marquardt optimization with mean square error as the performance function. This is a very robust method, which minimizes a linear combination of squared errors and weights. At the end of the training the resulting network has good generalization qualities. We describe the outcomes of our studies in the next section.

3. LOCAL RESULTS

3.1 Infusing CCNY LIDAR PBL Information

While strong correlations can be found connecting AOD to PM2.5 in the North East, the PBL height information can be expected to provide significant improvement. We now want to directly study the impact of adding the PBL in the neural network input. We use the 1064 nm channel measurement to derive the PBL height since this channel is less sensitive to molecular backscatter and attenuation, which means the lidar signal, is mostly contributed by aerosols. The PBL typically contains greater aerosol concentration than the overlying troposphere and hence has larger backscatter. The lidar data are processed using the Wavelet transform for PBL height calculations [26,27]. Note that the lidar can see aloft plumes and clouds allowing us to filter out these poor cases [28]. Fig. 4 show examples of the daily variability of the PBL height. During the cold months (Fig. 4(a)) PBL heights are small (HPBL< 1.2 km) in comparison to the heights observed during the summer (Fig. 4(b)). Moreover, the gradual increase of the planetary boundary layer observed in the warm months also indicates an unstable although well mixed nature. Finally, the use of lidar provides an additional quality control by removing cases with significant upper atmospheric plume activity (see Fig. 4(b)).

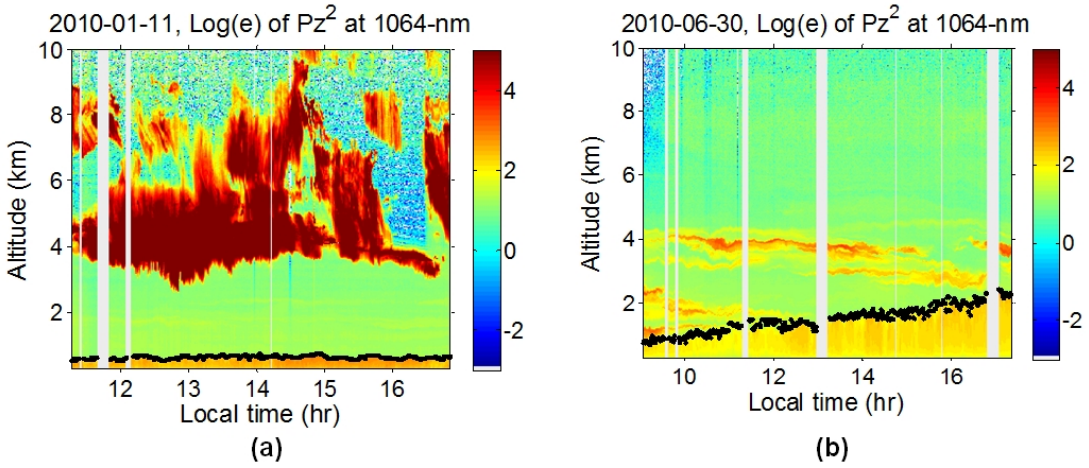

Fig. 4. CCNY Lidar images at 1064nm-channel and PBL-height (black dots): (a) January 11, 2010 and (b) June 30, 2010.

To further explore the role of PBL affecting PM2.5- AOD relationship we explore both the monthly and diurnal variations of PBL height. Fig. 5 shows the hourly and monthly variation of the PBL height during 2010, with higher values seen late in the afternoon during the summer (HPBL > 1.5 km). Smaller values (HPBL < 1.2 km) are recorded during most of the winter and fall. Again, more stable heights are retrieved when temperature is low in comparison to the diurnal cycle observed with high temperatures. Since most aerosols reside in the planetary boundary layer, the PBL height variability should clearly contribute in the analysis of the PM2.5–AOD relationship in order to improve model estimation.

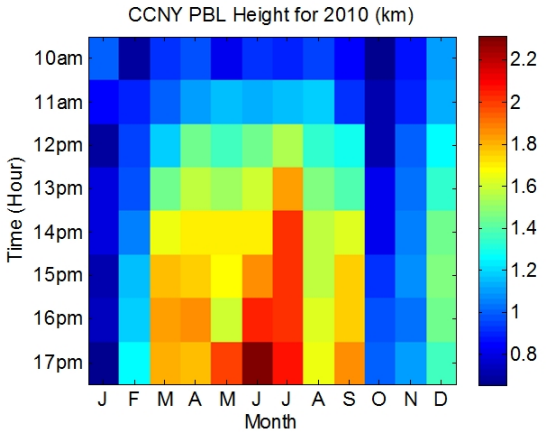

Fig. 5. Hourly-average PBL heights derived from lidar measurements in 2010.

3.1.1 Relevance of PBL information

In addition to supplying PBL height information, we explored the potential effect of meteorological variables such as temperature, relative humidity and wind speed in combination with the AOD as regressors to the neural network input. These case studies helped to discover the relevance of different variables by checking the efficiency of the neural network in estimating fine particulate matter. As shown in Fig. 6, yearly comparisons show that the planetary boundary layer height information resulted in the highest correlation (R~0.7) when paired with total AOD as inputs for the neural network over other parameters such as temperature, relative humidity and wind speed. Besides PBL information, the addition of temperature data also produced a considerable improvement in correlation (R~0.6). This is expected since the PBL height is correlated at some level with the surface temperature, as stronger convection and deeper PBL layers are associated with higher temperatures. All other variables show nearly no effect on the PM2.5-AOD relationship (R~0.5).

3.1.2 Seasonality

In estimating PM2.5, additional underlying seasonal changes can be expected to occur due to modifications of their environment, which would vary with the season. Fig. 7 illustrates this variation when applying total AOD and PBL height information together with month as inputs for our neural network during different seasons. As expected due to the well-mixed PBL development in an urban environment, summer returned the highest performance (R~0.94) while fall and spring performed similarly (R~0.77). Surprisingly, winter seems to shows a better performance (R~0.87) but it must be remembered that lidar measurements rely on human supervision so that measurements in winter are very scarce and it may be expected that with such scarce data, good data matchups may be made with the fewest number of cases for comparisons. The benefits of seasonal filtering, when compared to yearly results is most dominant in the summer which is expected since the PBL development and mixing is strongest during urban summer conditions. It should also be borne in mind that summer is

when the highest PM2.5 cases occur in general and where exceedance observations are most critical. This makes summer evaluation particularly useful.

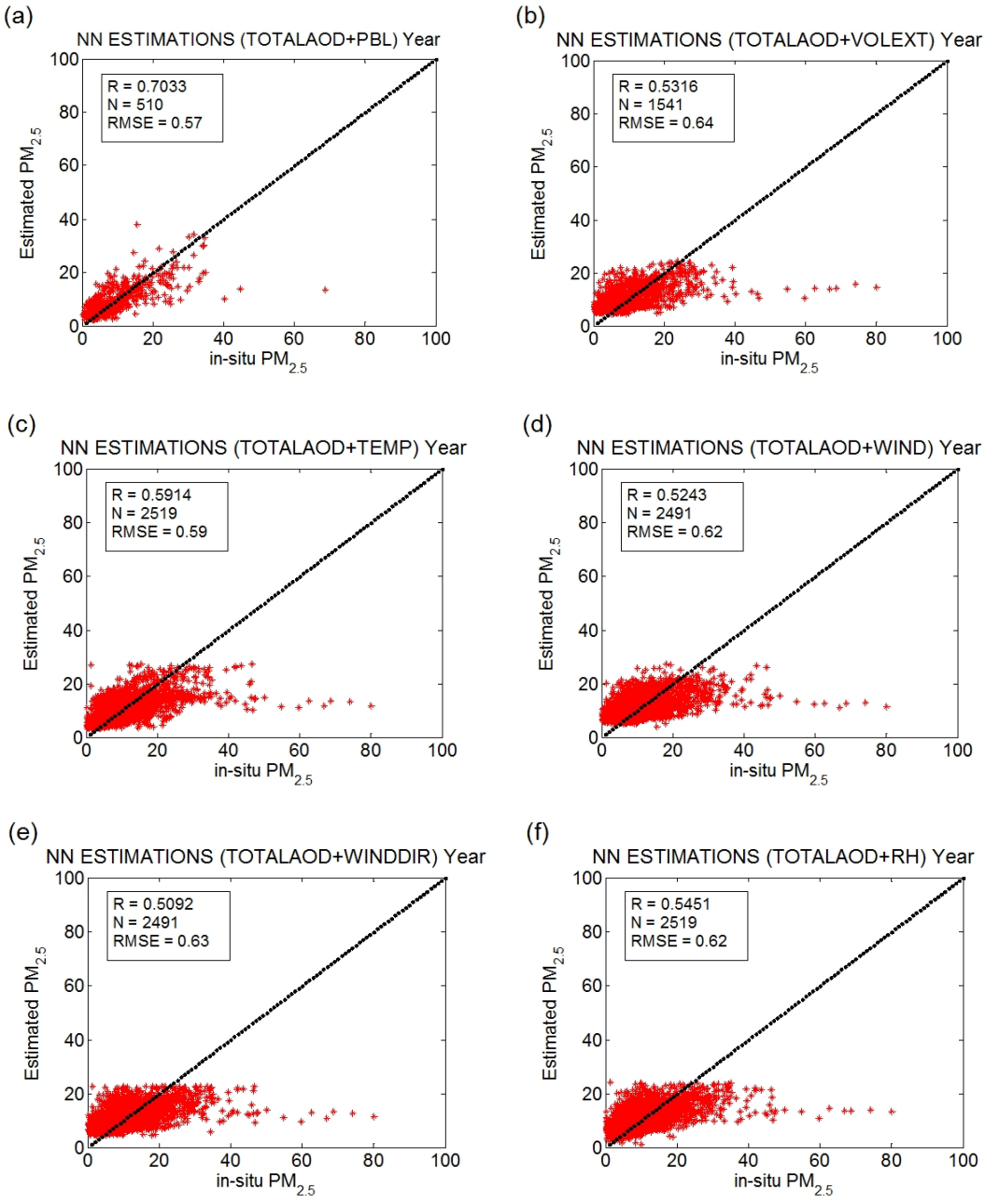

Fig. 6. Neural network results using additional parameters showing general improvements when other variables are considered in the PM2.5-AOD relationship

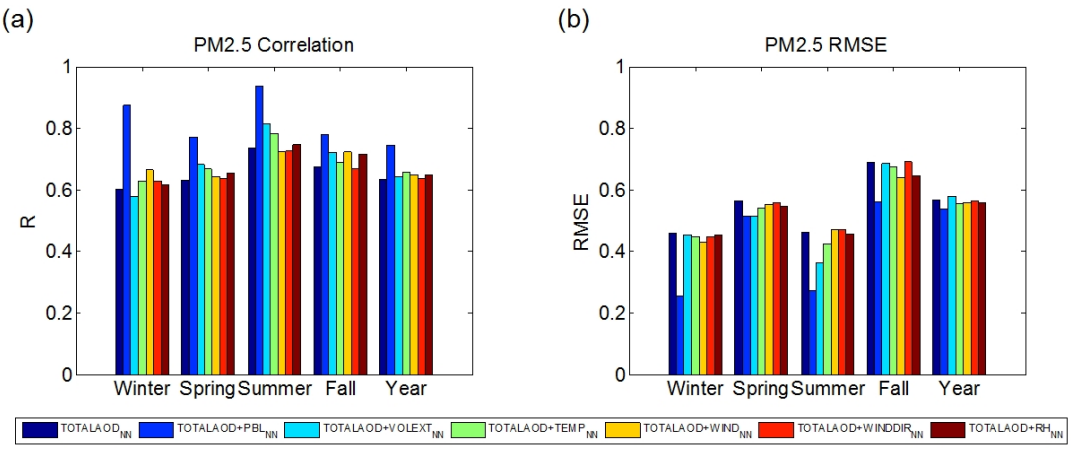

Fig. 7. (a) Seasonal correlation and (b) root-mean-square-error

While breaking up the data directly and using different neural networks for different cases may be optimal, it may be useful to explore the use of adding seasonal data directly into the NN training stream. To accomplish this we added month information as an extra input to our neural network and compared to our previous results. Using total AOD plus month as inputs, our correlation value returned a notable improvement (R~0.64). Table 1 show that addition of the month label consistently and significantly improves the correlation, confirming the PM2.5-AOD dependency on seasonality

Table 1. Comparisons between neural networks without/with month information added. The correlation (R) values are shown in the corresponding cells

Seasonal information	AOD only	AOD+ PBL	AOD+ Vol/Ext	AOD+ Temp	AOD+ Wind	AOD+ Wind-dir	AOD+ RH
None	0.5099	0.7033	0.5316	0.5914	0.5243	0.5092	0.5451
Month	0.6354	0.7451	0.6443	0.6573	0.6480	0.6380	0.6486

3.2 Assessment of Model PBL

Since our goal is to estimate PM2.5 on a bigger scale, our variables need to be available for a larger domain. Satellite AOD does provide spatial coverage at some extent with the combination of both Terra and Aqua. However, ground lidars are clearly not helpful and the space based CALIPSO lidar swath is so narrow that the orbit revisit time of 16 days is also not accommodating enough. Clearly, to compensate for the lack of ground instrumentation, the use of model data is essential for our regional application.

To evaluate the efficacy of the model in estimating PBL height, we compare it against CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations) derived PBL height. The Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) is the primary instrument on CALIPSO, which is part of the NASA A-train constellation of satellites [29]. Fig. 8 illustrates different features extracted from CALIPSO measurements along the track, including PBL height at different latitudes between 19:45 and 19:50UTC on August 8, 2007. In making comparisons, we filter matchups so that we use only single layer data in

order to assure no aloft layers or clouds are included which degrades both the lidar retrieval product as well as model meteorology effects.

Table 2 illustrates comparisons between CALIPSO and WRF PBL height. These comparisons were performed taking CCNY as center with a filter radius of 1 degree. From this, a reasonable agreement can be observed between the model and the instrument for the summer season which validates the use of WRF PBL into the neural network approach. Reasonable agreement can likely be obtained for the other seasons if the sample sizes were increased.

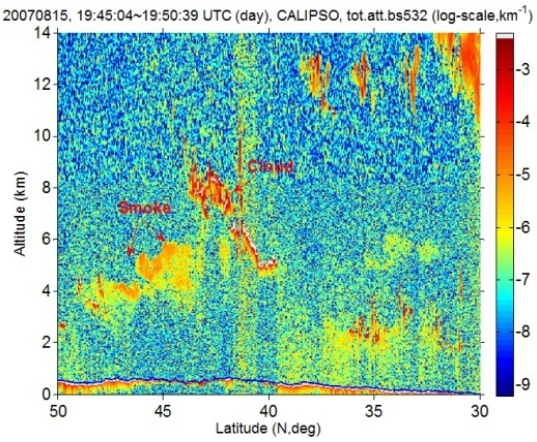

Fig. 8. CALIPSO derived PBL height on August 8, 2007. Smoke plumes and high cirrus clouds are also shown.

Table 2. Comparisons between CALIPSO and WRF derived PBL heights (around CCNY)

Radius	Winter		Spring		Summer		Fall	
	R	N	R	N	R	N	R	N
1.0 degree	0.37	75	0.41	24	0.63	49	0.44	35

As can be expected, the CALIPSO matchups are instructive in showing the seasonal improvement of summer in comparison to other seasons. This is at least partly due to the stronger aerosol layers making the CALIPSO measurements more accurate so this may not give a complete performance of the WRF model alone. On the other hand, when matching against the ground based lidar, even better matchups are seen since the ground based system is expected to better filter cases with higher accuracy and work better with weaker PBL aerosol layers (Fig. 9) [30].

Finally, we have made a multiyear analysis of summer PBL heights from CALIPSO which can be compared to existing climate model PBL forecast models [31,32]. It is clear that strong spatial correlations exist illustrating the applicability of model based PBL height directly into any PM2.5 estimator (Fig. 10). Comparisons for other seasons are much less accurate with significiant noise added to the CALIPSO retreivals.

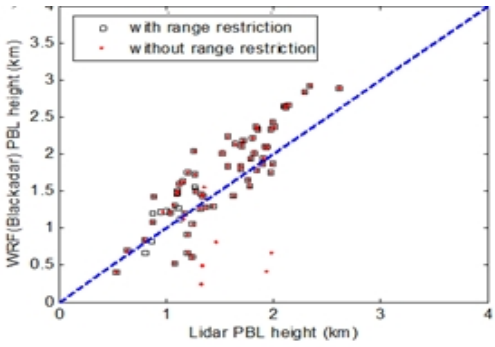

Fig. 9. CCNY lidar matchups against WRF model for summer 2007.

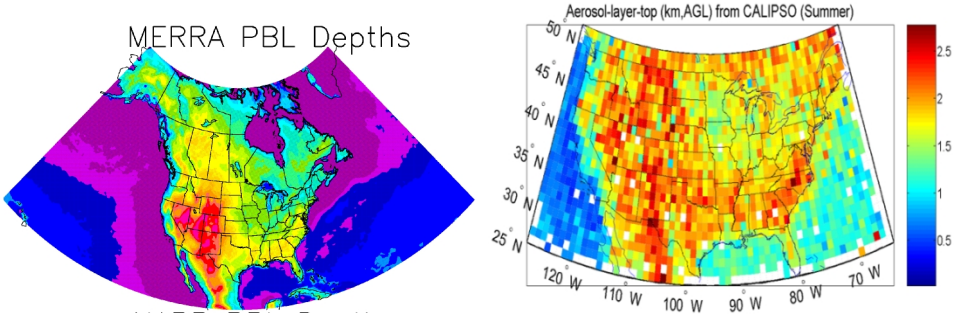

Fig. 10. (a) Summer climate model PBL Heights (b) CALIPSO PBL height retrievals

4. REGIONAL NEURAL NETWORK RESULTS

From our previous results, we've seen the importance of including PBL information as well as the seasonal factor in our neural network approach. Next, our regional experiment uses MODIS AOD and WRF PBL daily data to estimate 24-hour in-situ PM2.5 measurements which we compare against the performance from simple linear regressions, a neural network using daily MODIS AOD alone, and the GEOS-CHEM estimated PM2.5. Unlike the MODIS AOD daily averages, we only average the WRF PBL height data from 17:00 UTC to 20:00 UTC as a proxy for daily average since the WRF model performs poorly during night hours and late afternoon PBL displays a better representation of the actual PBL behavior. In this regional case, our neural network is mostly founded on collecting data from neighbor-stations based on a distance radius from a choosen center station. Besides, we include the month as a direct input since our local CCNY tests have confirmed the importance of adding this information for fine PM predictions. It should be pointed out that we have applied the GEOS-CHEM PM2.5 product quality flag keeping all points whose uncertainty is 20% of the mean. The results are shown in Fig. 11. The main observations are the significant improvements in the NN results with less cases of significant overestimation which appears in the GOES-CHEM case. We do see bias in the NN case at high PM2.5 which may be due to an unevenness in the frequency of high PM2.5 events and should be further explored. In Fig. 12(a-e), we demonstrate the procedure on a wider spatial domain to illustrate that the general procedures can be implemented on any scale. However, Fig. 12(f) is displayed on a finer domain surrounding NY State to make the comparisons with the NY state AIRNow stations (see Fig. 2) more visible. More details are provided in section 4.1.

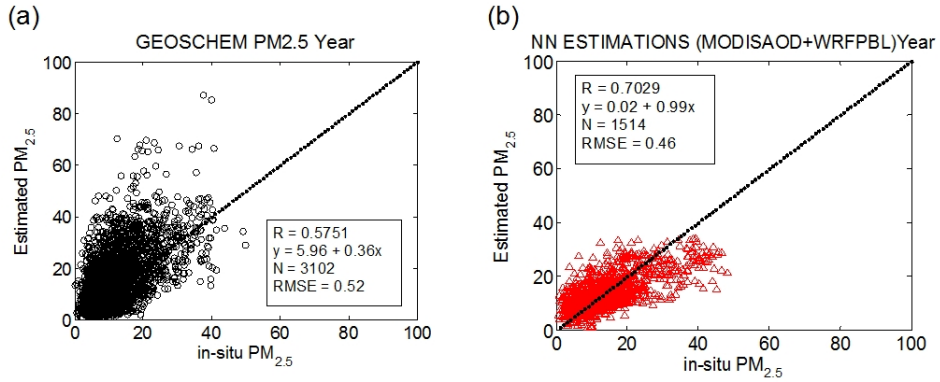

Fig. 11. Scatter plot showing the predicted output against in-situ ground station data a) Operational GEOS_CHEM results. b) NN approach including month input. The increased correlation and the decrease in RMSE value show better performance for the NN estimations

4.1 Implementing NN for PM2.5 Mapping

In order to estimate fine PM to conform to EPA standards, we use daily AOD retrievals together with WRF PBL measurements to illustrate our neural network performance as a spatial map. However, even in a "good" case, cloud cover can significantly reduce spatial coverage. Therefore, we utilize a simple preprocessing approach where we merge low resolution and high resolution data together with iterative inverse distance weighting.

In particular, we iteratively apply inverse distance weighted (IDW) averages of the data as defined in (1), to improve spatial coverage while using a 0.1 degree radial domain.

$$AOD_{IDW} = \frac{\sum_{i=1}^{n}(\frac{AOD}{d^2})_i}{\sum_{i=1}^{n}(\frac{1}{d^2})_i}, \quad (1)$$

As an example, we plot the cascade of intermediate results and final product in Fig. 12. for July 18, 2006. Fig. 12(a-b) show both the daily average AOD and the interpolated-spatial-coverage daily average AOD. For consistency, we used the same IDW approach to improve the spatial resolution in WRF PBL data although clearly data gaps are not an issue. Resulting images are shown in Fig. 12(c-d).

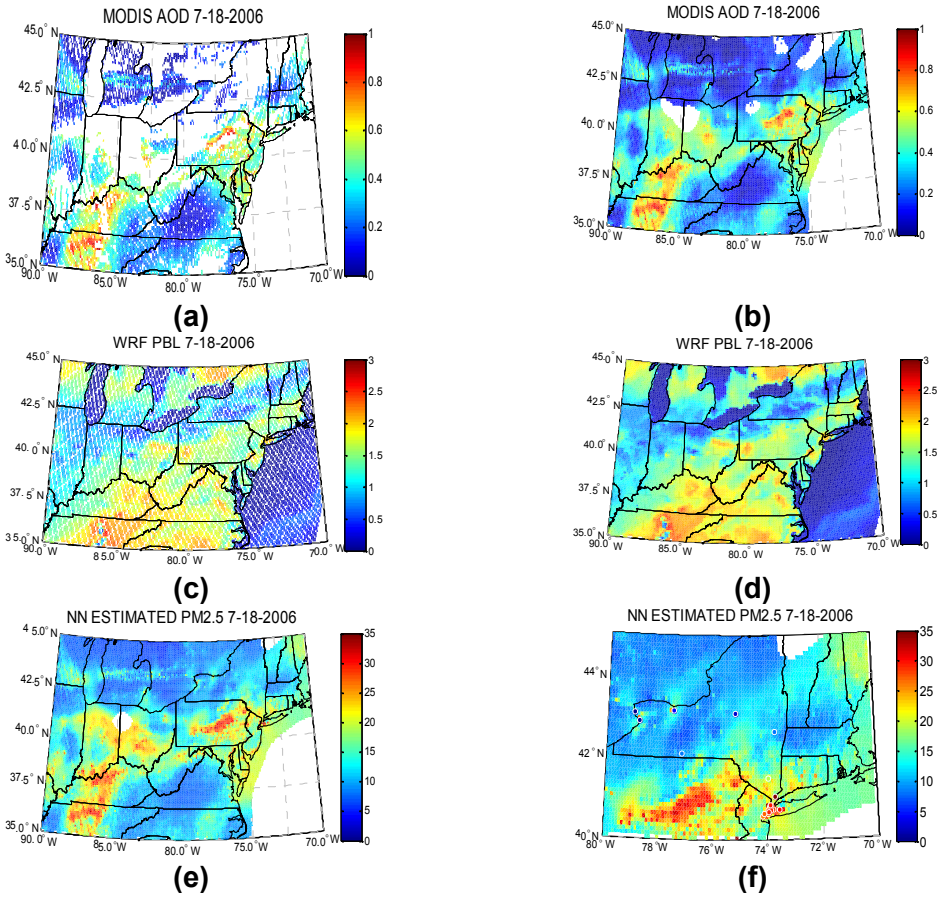

Fig. 12. (Left)Daily average and (Right) improved spatial coverage based on IDW for July 18, 2006. (a-b) MODIS AOD daily average, (c-d) WRF PBL daily average, (e-f) PM2.5 daily average estimations based on the neural network approach

Fig. 12(e-f) show the regional PM2.5 map obtained from our neural network together with a close up of the NY state region plus the station readings in that particulate date. We see a good agreement between station and estimations data with low PM2.5 values observed in the non-urban region while high fine PM values are observed in the metropolitan area.

Finally, for this day, Fig. 13 shows the scatter plot between ground measurements and estimations at the corresponding stations available in this particulate date. In general, neural network estimations are underestimated in comparison to site data; however, good correlation and RMSE values are obtained for the specific date (R~0.79, RMSE~0.47).

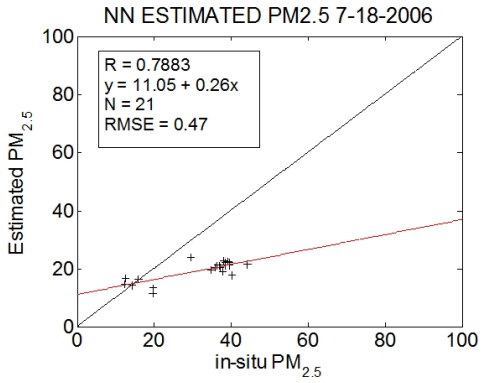

Fig. 13. Fine PM comparisons between ground measurements and neural network estimations on July 18, 2006

5. ASSESSMENT USING SATELLITE REMOTE SENSING VARIABLES

In the previous section we have demonstrated that underlying non-linear relationships between the AOD and PM2.5 can be partially reconstructed by applying neural network methods. In particular, the addition of PBL height and month inputs increased the performance of the neural network.

In this section we focus entirely on satellite observations, and study their potential influence on PM2.5 estimation. However, before we can start training the neural network, a careful choice of variables should be made. One of the ways would be to try a brute force search method, where all the possible combination of variables are fed into the neural network and the relevant set of variables are decided based upon the best correlated results between the prediction and the observation [33]. The brute force method is computationally intensive, and can be prohibitory for many variables case. Here, we successfully reduced the number of inputs by removing the correlated variables since the presence of redundant information results in unpretentious robustness during the testing phase. Fig. 14 shows the correlation coefficient between the candidate variables. In

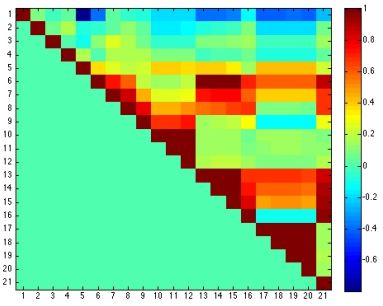

Fig. 14. The correlation map between 21 satellite variables. We remove the variables which are highly correlated (r>0.5). See text for details.

particular, we eliminated variables with cross-correlation coefficient greater than 0.5. This allowed us to significantly reduce the dimensionality of the input to the neural network. For example, “Corrected Optical Depth Land” at different wavelengths are highly correlated with each other and we could use one of them as the proxy.

Our initial regressor set started with the set of variables listed in Table 3 as possible candidates. The selections were made based upon the availability, and distribution spread of the variables. After removing the redundant variables (r>0.5), the number of variables were significantly reduced, which is shown in Table 3 with hash tags (#).

Table 3. The set of input variables are shown. The variables with hashtags (#) are selected for NN training after removing the correlated variables

Candidate satellite variables	Wavelength (um)
Solar_Zenith#	--
Solar_Azimuth#	--
Sensor_Zenith#	--
Sensor_Azimuth#	--
Scattering_Angle#	--
Optical_Depth_Land_And_Ocean#	0.55
Mean_Reflectance_Land_All	0.47
Mean_Reflectance_Land_All	0.66
Mean_Reflectance_Land_All#	2.1
Surface_Reflectance_Land	0.47
Surface_Reflectance_Land	0.66
Surface_Reflectance_Land	2.13
Corrected_Optical_Depth_Land	0.47
Corrected_Optical_Depth_Land	0.55
Corrected_Optical_Depth_Land	0.66
Corrected_Optical_Depth_Land_wav2p1	2.13
Optical_Depth_Small_Land	0.47
Optical_Depth_Small_Land	0.55
Optical_Depth_Small_Land	0.66
Optical_Depth_Small_Land	2.13
Mass_Concentration_Land	--

With the seven input variables as the regressors, we used the Bayesian regulation backpropagation method with 20 hidden nodes. Fig. 15 shows the result of the multivariate neural network training with only the 7 satellite variables taken as the input and the ground station PM data as the target.

In particular, we find that the multivariate satellite model using only the satellite remote sensing input variables are found to have the highest correlation with the target in comparison to the other methods. In fact, most of the biases seen in the PBL method seem to be diminished.

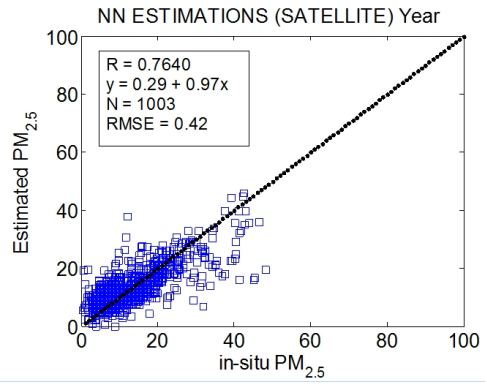

Fig. 15. Regression plot of satellite multivariate NN.

This clearly motivates further investigation of infusing surface as well as PBL height meteorological inputs. In addition, land surface classification is an important underlying indicator [34-36], and needs to be included. This can theoretically be done either using a modified vegetation index [37,38] or by direct land classification from USGS.

6. CONCLUSIONS

Our main focus was to demonstrate that regionally trained NN approaches are more accurate with less bias that operational approaches which do not directly input regional information. Before exploring the regional satellite based approach, we used combined active/passive radiometer and lidar measurements to assess PM2.5 estimation. First, we established that adding lidar derived PBL to total AOD is the most important "meteorological factor" that must be accounted for which is particularly reasonable for urban conditions when convective mixing is expected to be magnified. In addition, we also investigated the importance of seasonality into the PM2.5-AOD relationship because PBL and other factors, which affect physical and chemical properties of aerosol, also depend on seasons and found that separate seasonal training can provide significant improvements. In addition, using the month of the year as an additional training factor ('regressor') showed significant improvement.

The success of the local experiments motivated a study of combining satellite AOD and WRF PBL height which was methodically compared against GEOS-CHEM estimated PM2.5 where seasonal factors were integrated into the training. The results indicate that the regionally trained NN performs significantly better with much less over-bias at low PM2.5 values. With this approach, we developed daily PM2.5 maps using high resolution AOD and PBL grids for the region. Since the spatial coverage was still sparse, we applied an inverse distance weighting (IDW) approach and obtained better spatial coverage. The resulting maps are in good agreement with station data.

Finally, we also explored the potential of adding geometric and land surface satellite variables as additional regressors to the neural network. This is reasonable since it is generally understood that satellite AOD retrieval biases are connected to geometric and land surface factors. The results show better correlation and even lower biases at high PM2.5. Presently, it is quite remarkable that the satellite regressors when optimized seem to perform better than the use of satellite AOD and WRF PBL. However, it should be pointed out that the use of model PBL has significant issues such as inherent errors within the models as well as the fact that only in convectively mixed systems is the aerosol well mixed and trapped within the PBL. Still, additional research is in progress to include WRF PBL as well as temperature and relative humidity in combination with the satellite regressors. Further extensions to include neighboring states are also being investigated to improve the robustness of the approach.

ACKNOWLEDGMENTS

This project was made possible by the National Oceanic and Atmospheric Administration, Office of Education Educational Partnership Program award NA11SEC4810004 as well as NYSERDA under grant # 22885. Its contents are solely the responsibility of the award recipient and do not necessarily represent the official views of the U.S. Department of Commerce, National Oceanic and Atmospheric Administration (NOAA) or New York State Energy Research and Development Authority (NYSERDA).

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Pope CA, III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. J. of the American Medical Association. 2002;287(9):1132−1141.
2. U. S. Environmental Protection Agency. Air quality criteria for particulate matter. EPA/600/P-99/002aF, Research Triangle Park, N. C; 2004.
3. Hoff RM. Christopher SA. Remote sensing of particulate pollution from space: have we reached the promised land? J. Air. & Waste Manage. Assoc. 2009;59(6):645-675.
4. Zhang H, Hoff RM, Engel-Cox JA. The relation between Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth and PM2.5 over the United States: a geographical comparison by EPA regions. J. Air & Waste Manage. Assoc. 2009;59:1358-1369.
5. Tsai TC, et al. Analysis of the relationship between MODIS aerosol optical depth and particulate matter from 2006 to 2008. Atmos. Environ. 2009;45(27):4777-4788.
6. Boyouk N, Leon JF, Delbarre H, Podvin T, Deroo C. Impact of the mixing boundary layer on the relationship between PM2.5 and aerosol optical thickness. Atmos. Environ. 2010;44(2):271-277.
7. Tian J, Chen D. A semi-empirical model for predicting hourly ground-level fine particle matter (PM2.5) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements. Rem. Sens. Environ. 2010;114:221-229.
8. Gupta P, Christopher SA. An evaluation of Terra-MODIS sampling for monthly and annual particulate matter air quality assessment over the Southeastern United States. Atmos. Environ. 2008;42(26):6465.
9. Schaap M, Apituley A, Timmermans RMA., Koelemeijer RBA, de Leeuw G. Exploring the relation between aerosol optical depth and PM2.5 at Cabauw, the Netherlands. Atmos. Chem. Phys. 2008;9:909–925.
10. Hänel G. The properties of atmospheric aerosol particles as functions of the relative humidity at thermodynamic equilibrium with surrounding moist air. Adv. Geophys. 1976;19:73–188.
11. Liu Y, Sarnat JA, Kilaru V, Jacob DJ, Koutrakis P. Estimating Ground-Level PM2.5 in the Eastern United States Using Satellite Remote Sensing. Environ. Sci. Technol. *2005;39*(9):3269-3278.
12. van Donkelaar A*.,* Martin RV, Park RJ. Estimating ground-level PM2.5 using aerosol optical depth determined from satellite remote sensing. J. Geophys. Res. *2006;*111:*D21201.*
13. Yu SR, Mathur K, Schere D, Kang J, Pleim J, Young D, Tong G, Pouliot SA, McKeen, Rao ST. 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.
14. van Donkelaar A*,* Martin RV, Pasch *A*N, Szykman JJ, Zhang L, Wang YX, Chen D. Improving the accuracy of daily satellite-derived ground-level fine aerosol concentration estimates for North America. Environ. Science & Technology. 2012;46(21):11971-11978**.**
15. Hu X, Nielsen-Gammon JW, Zhang F. Evaluation of Three Planetary Boundary Layer Schemes in the WRF Model, J. Appl. Meteor. Climatol. 2010;49:1831–1844.

16. Hogrefe C, Doraiswamy P, Colle B, Demerjian K, Hao W, Beauharnois M, Erickson M, Souders M, Ku JY. Effects of Grid Resolution and Perturbations in Meteorology and Emissions on Air Quality Simulations Over the Greater New York City Region, CMAS Conference Baltimore (2011)
17. Holben BN, et al. AERONET - A federated instrument network and data archive for aerosol characterization. Rem. Sens. Environ. 1998;66:1-16.
18. O'Neill NT, Eck TF, Smirnov A, Holben BN, Thulasiraman S. pectral Discrimination of Coarse and Fine Mode Optical Depth. J. Geophys. Res. 2003;108(D17):4559.
19. Dubovik O, Smirnov A, Holben BN, et al. Accuracy assessments of aerosol optical properties retrieved from AERONET Sun and sky radiance measurements. J. Geophys. Res. 2000;105:9791–9806.
20. Wu Y, Chaw S, Gross B, Moshary F, Ahmed S. Low and optically thin cloud measurements using a Raman-Mie lidar. Appl. Opt. 2009;48:1218e1227.
21. Yonghua Wu, Chuen Meei Gan, Lina Cordero, Barry Gross, Fred Moshary, Sam Ahmed. Calibration of the 1064 nm lidar channel using water phase and cirrus clouds, Appl. Opt. 2011;50:3987-3999.
22. Gupta P, Christopher SA. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. J. Geophys. Res. *2009;114:D14205.*
23. Bishop C. Neural Networks for Pattern Recognition Oxford Univ. Press, 1996.
24. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.
25. MacKay DJC. Bayesian interpolation, Neural Computation, 1992;4(3):415–447.
26. Davis KJ, Gamage N, Hagelberg CR, Kiemle C, Lenschow DH, Sullivan PP. An objective method for deriving atmospheric structure from airborne lidar observations. J. Atmos. and Oceanic Technol. 2000;17:1455–1468.
27. Brooks IM. Finding boundarylayer top: Application of a wavelet covariance transform to Lidar Backscatter Profiles. J. Atmos. and Oceanic Technol. 2003;20:1092–1105.
28. Lolli S, Delgado R, Compton J, Hoff R. Planetary boundary layer height retrieval at UMBC in the frame of NOAA/ARL campaign, Proc. SPIE, 2011;8182:81820R.
29. Winker DM, Vaughan MA, Omar A, et al. Overview of the CALIPSO mission and CALIOP data processing algorithms. J. Atmos. Oceanic Tech. 2009;26(11):2310-2323.
30. Gan CM, Wu Y, Madhavan BL, Gross B, Moshary F. Application of active optical sensors to probe the vertical structure of the urban boundary layer and assess anomalies in air quality model PM2.5 forecasts. Atmospheric Environment. 2011;45(37):6613-6621.
31. McGrath-Spangler, Erica L. Importance of boundary layer entrainment for surface fluxes over land, Ph.D. thesis, Colorado State Univ, 2011.
32. Jordan NS, Hoff RM, Bacmeister JT. Validation of Goddard Earth Observing System version 5 MERRA planetary boundary layer heights using CALIPSO. J. Geophys. Res. 2010;115:D24218.
33. Malakar NK, Lary DJ, Moore A, Gencaga D, Roscoe B, Albayrak A, Wei J. Estimation and bias correction of aerosol abundance using data-driven machine learning and remote sensing. Intelligent Data Understanding (CIDU), 2012 Conference on, 2012;24(300;24-26.
34. Remer LA, Kaufman YJ, Tanre D, Mattoo S, Chu DA, Martins JV, Li RR, Ichoku C, Levy RC, Kleidman RG, Eck TF, Vermote E, Holben BN. The MODIS aerosol algorithm, products, and validation. J. Atmos. Sci. 2005;62(4):947-973.

35. Levy RC, Remer LA, Mattoo S, Vermote EF, Kaufman YJ. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance. J. Geophys. Res. Atmos. 2007;112(D13):D13211.
36. Lary DJ, Remer LA, MacNeill D, Roscoe B, Paradise S. Machine Learning and Bias Correction of MODIS Aerosol Optical Depth. Geoscience and Remote Sensing Letters, IEEE. 2009;6(4):694-698.
37. Oo MM, Gross B. Measurement of Aerosol Properties over Urban Environments from Satellite Remote Sensing, Advances of Environmental Remote Sensing to Monitor Global Changes, CRC Press Ni-Ban Chang Ed. Chapter 15, 2012.
38. Oo MM, Hernandez E, Jerg M, Gross B, Moshary F, Ahmed SA. Improved MODIS Aerosol Retrieval Using Modified VIS/MIR Surface Albedo Ratio Over Urban Scenes, IEEE TGRS. 2010;48:983-1000.

APPENDIX 1

Table A.1 - New York state stations

City	Latitude	Longitude	AQS Number	DEM (m)	Urban/non-urban
PS 274	40.69454	-73.92769	360470118	14	1
IS 293	40.68419	-73.99298	360470121	14	1
IS 74	40.81551	-73.88553	360050112	14	1
PS 154	40.80833	-73.92612	360050113	14	1
PS 314	40.64182	-74.01871	360470052	11	1
Manhattanville P.O.	40.81133	-73.95321	360610119	14	1
Maspeth	40.72698	-73.89313	360810120	14	1
PS 44	40.63137	-74.15732	360850114	11	1
IS 52	40.81620	-73.90200	360050110	14	1
White Plains	41.05192	-73.76366	361192004	108	0
Park Row	40.71160	-74.00540	360610125	10	1
Eisenhower Park	40.74316	-73.58549	360590005	33	0
IS 143	40.84888	-73.93059	360610115	40	1
Division Street	40.71436	-73.99518	360610134	14	1
Fresh Kills West	40.58027	-74.19832	360850111	14	1
Queens College	40.73614	-73.82153	360810124	20	1
CCNY	40.81976	-73.94825	360610135	14	1
PS 19	40.73000	-73.98446	360610128	14	1
Newburgh	41.49916	-74.00885	360710002	180	0
Albany County HD	42.64225	-73.75464	360010005	58	0
Buffalo	42.87691	-78.80981	360290005	198	0
East Syracuse	43.05235	-76.05921	360671015	140	0
Lackawanna	42.8273	-78.84984	360291007	75	0
Loudonville	42.68075	-73.75733	360010012	68	0
Niagara Falls	43.08218	-79.00106	360632008	173	0
Pinnacle	42.09142	-77.20978	361010003	444	0
Rochester	43.14618	-77.54817	360551007	323	0
Rockland County	41.18208	-74.02819	360870005	245	0
Tonawanda II	42.99813	-78.89926	360291014	181	0
Utica	43.09892	-75.22506	360652001	233	0
Westfield	42.29071	-79.58961	360130011	406	0
Whiteface Lodge	44.39308	-73.85890	360310003	613	0
Babylon	40.74529	-73.41919	361030002	34	0
Hempstead	40.63100	-73.73390	360590008	2	0
JHS 126	40.71961	-73.94771	360470122	14	1
JHS 45	40.79970	-73.93432	360610079	14	1
Mamaroneck	40.93149	-73.76575	361191002	18	0
Morrisania II	40.83606	-73.92009	360050080	40	1
NY Botanical Garden	40.86790	-73.87809	360050133	40	1
Port Richmond	40.63307	-74.13719	360850055	11	1
Susan Wagner	40.59664	-74.12525	360850067	11	1

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=323&id=10&aid=2530