Estimation and Bias Correction of Aerosol Abundance using Data-driven Machine Learning and Remote Sensing

N. K. Malakar, D. J. Lary, A. Moore, D. Gencaga, B. Roscoe
University of Texas at Dallas,
Richardson, TX

A. Albayrak, and J. Wei
Goddard Earth Sciences DISC,
NASA Goddard Space Flight Center

Abstract

Air quality information is increasingly becoming a public health concern, since some of the aerosol particles pose harmful effects to peoples health. One widely available metric of aerosol abundance is the aerosol optical depth (AOD). The AOD is the integrated light extinction coefficient over a vertical atmospheric column of unit cross section, which represents the extent to which the aerosols in that vertical profile prevent the transmission of light by absorption or scattering. The comparison between the AOD measured from the ground-based Aerosol Robotic Network (AERONET) system and the satellite MODIS instruments at 550 nm shows that there is a bias between the two data products. We performed a comprehensive search exploring possible factors which may be contributing to the inter-instrumental bias between MODIS-Aqua land data set and AERONET. The analysis used several measured variables, including the MODIS AOD, as input in order to train a neural network in regression mode to predict the AERONET AOD values. This not only allowed us to obtain an estimate, but also allowed us to infer the optimal sets of variables that played an important role in the prediction. In addition, we applied machine learning to infer the global abundance of ground level PM2.5 from the AOD data and other ancillary satellite and meteorology products.
1 Introduction

Atmospheric aerosols are tiny particles (solid and liquid) suspended in the atmosphere. Inhalation of some aerosols produce harmful effects to people’s health. Moreover, atmospheric aerosols play an important role in understanding the global climate.

The aerosol optical depth (AOD), or optical thickness, is defined as the integrated extinction coefficient over a vertical column of unit cross section. The extinction coefficient is the fractional depletion of radiance per unit path length and represents how much the aerosols prevent the transmission of light by absorption and scattering.

In the past, much effort has been placed in observing aerosol characteristics, such as AOD, from space and ground-based instruments. The Moderate Resolution Imaging Spectroradiometer (MODIS), onboard the Terra and Aqua satellites, retrieve AOD using dark target methods in bands at 550, 670, 870, 1240, 1630, and 2130 nm over the ocean, and at 470, 550, and 670 nm over land [1, 2]. A global system of ground-based sun and sky scanning sun photometers, called the Aerosol Robotic Network (AERONET), also measure the AOD at various wavelengths (at 340, 380, 440, 500, 675, 870, and 1020 nm) [3]. AERONET measurements are taken every 15 min during daylight, and its level 2 quality control measurements assure AOD observations are accurate to within 0.01 for wavelengths of 440 nm and higher. AOD measurements from MODIS are available globally, whereas AERONET measurements are available only for land locations, some of which are coastal sites. In this paper, we analyzed MODIS-Aqua land data set.

Ideally, the measurements of AOD from these two instruments should match. However, biases do exist between AERONET and MODIS measurements. In this study we refer to the difference between the ground truth AERONET AOD observations at 500 nm and the remotely sensed MODIS AOD at 550 nm as the bias. Figure 1 shows that a significant number of points do not fall close to the 1:1 line. Although the high-AOD biases are larger, the small AOD biases are equally important. Our goal is to try and understand the factors that can delineate these extrema, and/or explain them statistically.
Fig. 1: A scatter diagram showing the comparison between the AOD from AERONET and MODIS instruments at 550 nm. A circular region has been highlighted to show a high bias regime.

2 Previous Studies

Previous MODIS aerosol validation studies compared the Aqua and Terra MODIS-retrieved AODs with the ground-based AERONET observations [4, 5, 2, 6].

From the studies of Normalized Difference Vegetation Index (NDVI), Brown et al. (2008) suggested that the surface type played a key role in explaining a significant fraction of the observed bias [7].

Lary et al. (2009) used machine-learning approaches to explore factors contributing to a persistent bias between AOD retrieved from MODIS and AERONET data [8]. Their work also suggested a link between the MODIS AOD bias and the surface type. The possible factors influencing the bias might be associated with the measurement conditions such as the solar and sensor zenith angles, the solar and sensor azimuth, scattering angles, and surface reflectivity at the various measured wavelengths, etc. In their study they explained the AOD bias between MODIS and AERONET by using the surface type, the solar zenith angle, the solar azimuth angle, the sensor zenith angle, the sensor azimuth angle, the scattering angle, and the reflectance at 550 nm as input variables to the neural network.

In this paper, we performed a comprehensive analysis for every possible combination of the variables as input to train the neural network in regression mode to predict the AOD values. We then compared how well the predictions
matched with the observed AOD values. As a result, we obtained the best set of variables explaining the bias in the MODIS (AOD) measurements.

3 Data

In this study we used data set derived from Multi-sensor Aerosol Products Sampling System (MAPSS). The MAPSS framework derives the data from multiple sources such as original MODIS and AERONET datasets, and provides a uniform and consistent sampling of aerosol products [9]. The data sampling technique used by MAPSS is extension of the sampling approach developed by Ichoku et al [10]. However, instead of using $50 \times 50$ km square grids, the method identifies the MODIS data pixels within 25 km of the AERONET sites. Readers interested in the details of the MAPSS are directed to [9]. In the present paper we present the analysis of only the MODIS-Aqua land data set.

4 Neural Network Regression Technique

Neural networks (NN) are biologically inspired algorithms used for classification or function approximation [11, 12, 13]. NNs are widely used in pattern recognition, machine learning and artificial intelligence. In addition, NNs have found many applications in other fields such as geoscience, remote sensing, oceanography, etc. Neural networks are also referred to as a multi-layer perceptron method because they may consist of multiple layers (e.g., input, hidden and output layers). Each neuron is connected to all other neurons in the adjacent layers. Each neuron then is assigned weights for each interconnection with all the other neurons. The output of the $k^{th}$ neuron can be written as the weighted sum of inputs

$$y_k = \varphi \left( \sum_{j=1}^{n} w_{kj}x_j \right),$$

(1)

where $\varphi$ is the transfer function, $w_{kj}$ represents the weight from unit $j$ to unit $k$, and $x_j$ represents the $n$ input variables to the neuron. During training, the NN weights are adjusted appropriately to learn the data. The learning and adjustments of the weights are inspired by the synaptic learning behavior of neurons.
Fig. 2: Supervised neural network technique.

For an observation data set with \( n \) input variables, say \( \{x_1, x_2, x_3, \ldots, x_n\} \), the observed output variable, AOD, is some function of these input variables. Our approach approximates the function by non-parametric, non-linear NNs. We selected supervised NN method, since a NN learns from its input parameters and is free from assumptions about its inputs. This allows us to explore various sets of inputs. The goal here is to train the NN against the AERONET AOD data as the target, as shown in the figure 2. The trained NN is then used to predict the AOD for the given set of variables.

Next, we applied a neural network regression technique to learn the inter-instrumental bias and seek the best set of variables contributing to the bias.

5 Search for optimal set of variables for bias reduction

We input, as regressors to the NN, the AOD at 550 nm along with 14 other variables that are listed below. For brevity, we have denoted the variables by the corresponding numbers in the following table.

1. Aerosol optical depth at 550 nm (AOD0550)
2. Aerosol optical depth at 470 nm (AOD0470)
3. Aerosol optical depth at 660 nm (AOD0660)
4. Mean reflectance at 470 nm (mref0470)
5. Mean reflectance at 550 nm (mref0550)
6. Surface reflectance at 660 nm (surfre0660)
7. Surface reflectance at 470 nm (surfre0470)
8. Surface reflectance at 2100 nm (surfre2100)
9. Cloud fraction from land aerosol cloud mask (cfrac)
10. Quality assurance (QAavg)
11. Solar zenith angle (SolarZenith)
12. Solar azimuth angle (SolarAzimuth)
13. Viewing zenith angle (SensorZenith)
14. Sensor azimuth angle (SensorAzimuth)
15. Scattering angle (ScatteringAngle)

The number of combinations for n variables, considering a set of k at a time, is given by the combination \( \binom{n}{k} \). Our AOD data set contains 15 measured variables, and we considered all the possibilities such as \( \binom{15}{15}, \binom{15}{14}, \binom{15}{13}, \ldots, \binom{15}{2} \). Thus, there are 32,781 possible combinations to be explored. So, we made this a search problem where the search is over the possible set of variables that best fit the observed data. At end, the non-relevant variables are absent from the best-fitting set of variables.

For each combination set, one at a time, we trained the NN with AERONET AOD as the target variable, and then predicted what this AERONET AOD is from the trained network. The NN algorithm used a feed-forward back propagation algorithm with a hidden layer having 200 nodes, as shown in figure 3. The training was done by the Levenberg-Marquardt algorithm with mean-squared error as the performance factor, provided by the Matlab NN toolbox. When training a neural network we randomly split the training data set into three portions, in the ratio of 80 : 10 : 10. The first 80% portion is used to train the NN weights using an iterative process. For each iteration, we evaluate the current root mean square (RMS) error of the neural network by using the second 10% portion of the data (this portion was not used in the training). We use the RMS error and how it changes with the training iterations (known as epochs) to determine the convergence of our training. When the training is complete, we use the final 10% of the data as the validation data set.

Since the neural network constructs a mapping between the set of input and output variables, the most relevant set of variables is the one that can
Fig. 3: Matlab’s neural network toolbox was used to train the neural nets. A screenshot of 12 variable training case is shown.

best reproduce the target data. We explored all combinations of variables, which provided the fit of the observed AERONET AOD data. The end product is the result of regression between the input variables to the NN and the observed AERONET AOD. This is a massive number crunching exercise. We automated the workflow for each combination by writing a job-parallel code.

6 Similarity measure between Predicted and Observed AOD

In order to quantify the agreement between the observed and predicted data, we used both the correlation coefficient appropriate for Gaussianly distributed variables and the mutual information appropriate for variables of arbitrary probability distribution. The predictions made by the most relevant set of variables show the highest correlation coefficient or highest mutual
information with the observed data. In the appendix, we show that Mutual Information (MI) is the more general case of correlation coefficient. For normal distribution, the expression of MI returns the correlation measure. So, it makes sense to use MI as the general measure of correlation between the observed and predicted set, as many of the variables are not normally distributed.

In the literature, there are several methods to estimate MI from data [14, 15, 16]. We applied the Variable Bin Width Histogram Approach [17, 18] to compute the normalized MI between the observed and predicted AOD. Higher values indicate better agreement between the observed and predicted set, and thus are the best indicators of the input variables needed to assess a relevant set of variables. Table-1 contains the MI for all sets in decreasing order.

We can construct the best regression fit of the MODIS parameters to predict the AERONET AOD when certain MODIS parameters are absent in the combination. Table 2 shows the absent variables from the combination. These absent variables include aerosol optical depth at 550 nm (AOD0550), AOD at 660 nm, cloud fraction from land aerosol cloud mask (cfrac), surface reflectance at 470 nm (surfre0470), surface reflectance at 660 nm (surfre0660),
sensor azimuth angle (sensor azimuth), solar zenith angle (solar zenith). The fact that the Neural Networks performed better in the absence of certain variables indicates that these variables contribute to the observed bias.

Therefore, the methodology of comprehensive search provides us insights into the factors explaining the bias between the MODIS AOD and AERONET AOD. We also obtain the best performing NN and used it to estimate the bias corrected AOD observations [19]. Figure 4 shows the bias-corrected AOD plot compared to the AERONET AOD. Clearly, the bias at the higher values of the AOD have been corrected, and the AOD values follow values close to 1 : 1 line. This best prediction was obtained by a set consisting of the following variables: AOD at 470 nm, and AOD at 660 nm, mean reflectance at 470 nm, and mean reflectance at 550 nm, surface reflectance at 660 nm, 470 nm, 2100 nm, cloud fraction, quality assurance values, solar zenith angle, solar azimuth angle, zenith angle, sensor azimuth angle and scattering angle.

### 7 Estimating global PM2.5 abundance

As an example of applications of the framework, we briefly present an application to infer the global abundance of the ground-level-distribution of particles with a diameter of 2.5 micrometers (PM 2.5) or less from the AOD data and other ancillary satellite and meteorology products.

The abundance of PM 2.5 at ground level is known to adversely impact...
Tab. 2: Absent variables in Table 1

<table>
<thead>
<tr>
<th>Row #</th>
<th>Absent variables in the set as shown in the Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AOD0650</td>
</tr>
<tr>
<td>2</td>
<td>AOD0660, cfac, SensorAzimuth</td>
</tr>
<tr>
<td>3</td>
<td>surf60470, cfac</td>
</tr>
<tr>
<td>4</td>
<td>AOD0660, surf60470</td>
</tr>
<tr>
<td>5</td>
<td>SolarZenith, SensorAzimuth</td>
</tr>
<tr>
<td>6</td>
<td>AOD0660, surf60660, cfac, SolarZenith</td>
</tr>
<tr>
<td>7</td>
<td>AOD0660, ScatteringAngle</td>
</tr>
<tr>
<td>8</td>
<td>AOD0550, cfac, ScatteringAngle</td>
</tr>
<tr>
<td>9</td>
<td>AOD0470, cfac</td>
</tr>
<tr>
<td>10</td>
<td>AOD0660, surf62100, cfac</td>
</tr>
</tbody>
</table>

public health. For example, it has serious impacts on people with heart diseases, asthma, etc. [20]. Ground monitoring stations are only available at certain locations, so we do not have in-situ observations of ground level PM 2.5 (GLP) for the whole planet. Moreover, it is not always possible to obtain the GLP in areas with rapidly increasing population, which frequently harbor newer locations of significant air pollution. However, with remotely sensed data, and application of machine-learning, we can provide a daily GLP data product for the entire globe. By constructing an automated workflow with large computing facilities, it is possible to examine the GLP for any location in the world and to analyze the trends in global air pollution.

Figure 5 shows the GLP for the continental USA generated from satellite data, weather analysis, and roadside GLP observations. Currently there is full coverage of ground level PM 2.5 in the US, and nearly-full global coverage. The estimation of GLP has important health applications. Since the GLP products can be used to construct a global air pollution map, one could build an air-quality advisory for personal digital assistant (PDA). Global estimation with continuous spatial and temporal coverage can be critically useful in making real-time health care decisions as well as public policy decisions for improving the GLP and other environmental conditions. This is work in progress.

8 Results, Discussions and Future Work

We have presented a scheme of comprehensive search technique, with plug and play in mind, so that the machine learning technique of choice can be
easily coupled to perform a job-parallel analysis of massive number crunching tasks.

Specifically, in this paper, we studied factors influencing the bias in the observed AOD values between the MODIS and AERONET instruments. We applied the Supervised Neural Network method in regression mode to train the NN with the AERONET data set as the target, and recomputed the prediction of AOD from the neural nets. We performed an exhaustive search for the possible combinations of input variables.

We used Mutual Information, as the measure of similarity between the observed and bias-corrected/predicted data, which removes any need to assume the variables are normally distributed.

Using the neural network method, we found that the best prediction of AOD was provided by the set consisting of the following variables as the regressors: AOD at 470 nm, 660 nm, Mean reflectance at 470 nm, 550 nm, Surface reflectance at 660 nm, 470 nm, 2100 nm, Cloud fraction, Quality assurance values, Solar zenith angle, solar azimuth angle, Zenith angle, Sensor azimuth angle and scattering angle. The best agreement between the observed and predicted AOD occurred when some of the variables were missing from the input combinations. Similarly, for various other combinations, the absence of one or couple of variables, such as the AOD at 660 nm, the cloud fraction from the land aerosol cloud mask, the surface reflectance at 470 nm and at 660 nm, the sensor azimuth angle, the solar zenith angle etc. seems to indicate that their presence contributes to the observed bias.

The method was also applied to estimate the ground level PM2.5, which are known to have adverse health effects. In this case, we used the AOD
measurements and ancillary information as input and trained the NN in regression mode to estimate the GLP. The global estimation with a continuous spatial and temporal coverage can critically help in making public policy decisions for improving the GLP environmental conditions and healthcare decisions.

Neural network is one of the methods that we focused in this paper. We are working on implementing various machine-learning techniques to estimate and identify the variables contributing to the bias correction of AOD. In the future work we will present cross examination of multiple machine learning techniques to explain the bias-correction using the framework developed in this paper.

9 Acknowledgements

The support from NASA ACCESS AeroStat project (NNX10AM94G) is gratefully acknowledged. Special thanks are due to Dr. Maksym Petrenko and Dr. Charles Ichoku for providing Multi-sensor Aerosol Products Sampling System (MAPSS) data sets [9]. We also thank the AERONET investigators, and the MODIS team for their wonderful work.

Appendix

Correlation coefficient (Pearson’s correlation) is a widely used measure of dependence between two variables, and represents the normalized measure of the strength of their linear relationships. The correlation coefficient $\rho_{X,Y}$ between two random variables $X$ and $Y$ with expected values $\mu_X$ and $\mu_Y$ and standard deviations $\sigma_X$ and $\sigma_Y$ is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$  \hspace{1cm} (2)$$

where, E is the expected value operator, cov means covariance, and, $\rho$ a widely used alternative notation for Pearson’s correlation.

The correlation coefficient is defined only if both of the standard deviations are finite and both of them are nonzero. The correlation coefficients range from -1 to 1. The correlation coefficient values close to 1, (or -1) suggest that there is a positive (or negative) linear relationship between the data columns, whereas the values close to or equal to 0 suggest there is no linear
relationship between the data columns. It can only be applied to the cases of linear relationship between two variables.

Mutual information quantifies the mutual dependence between two variables by taking into account all of the characteristics of the variables in the Probability Distribution Function (PDF). Mutual information (MI) is defined as follows in discrete form:

\[
I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)},
\]

which is a special case of a measure called Kullback-Leibler divergence [21, 22]. If \( X \) and \( Y \) are statistically independent, then

\[
p(x,y) = p(x)p(y).
\]

In this case, the mutual information becomes 0, showing independency. A proper mapping of the form

\[
\delta(X,Y) = \sqrt{1 - e^{-2I(X,Y)}}
\]

normalizes the measure of general correlation as depicted by the MI [23, 24, 25]. In the case when \( X \) and \( Y \) are normally distributed,

\[
(X,Y) \sim N(\mu, K)
\]

where, \( K = (\sigma^2, \rho \sigma^2; \rho \sigma^2, \sigma^2) \). The mutual information reduces to

\[
I(X,Y) = -\frac{1}{2} \log(1 - \rho^2).
\]

So, that,

\[
\delta(X,Y) = \sqrt{1 - e^{-2I(X,Y)}} = |\rho(X,Y)|.
\]

This relation shows the generality of the normalized correlation measure.

References


