Impacts of snow and cloud covers on satellite-derived PM2.5 levels

Jianzhao Bi, Emory University
Jessica H. Belle, Emory University
Yuije Wang, University of Maryland Baltimore
Alexei I. Lyapustin, Univ Maryland Baltimore Cty
Avani Wildani, Emory University
Yang Liu, Emory University

Journal Title: Remote Sensing of Environment
Volume: Volume 221
Publisher: Elsevier Science Inc. | 2019-02-01, Pages 665-674
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1016/j.rse.2018.12.002
Permanent URL: https://pid.emory.edu/ark:/25593/vmb2d


Copyright information:
© 2018 Elsevier Inc. All rights reserved.

This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Accessed December 31, 2023 6:27 AM EST
Impacts of snow and cloud covers on satellite-derived PM$_{2.5}$ levels

Jianzhao Bi$^a$, Jessica H. Belle$^a$, Yujie Wang$^{b,c}$, Alexei I. Lyapustin$^{b,c}$, Avani Wildani$^d$, Yang Liu$^{a,*}$

$^a$Department of Environmental Health, Emory University, Rollins School of Public Health, Atlanta, GA, USA

$^b$Goddard Earth Sciences and Technology Center, University of Maryland Baltimore County, Baltimore, MD, USA

$^c$NASA Goddard Space Flight Center, Greenbelt, MD, USA

$^d$Department of Computer Science, Emory University, Atlanta, GA, USA

Abstract

Satellite aerosol optical depth (AOD) has been widely employed to evaluate ground fine particle (PM$_{2.5}$) levels, whereas snow/cloud covers often lead to a large proportion of non-random missing AOD values. As a result, the fully covered and unbiased PM$_{2.5}$ estimates will be hard to generate. Among the current approaches to deal with the data gap issue, few have considered the cloud-AOD relationship and none of them have considered the snow-AOD relationship. This study examined the impacts of snow and cloud covers on AOD and PM$_{2.5}$ and made full-coverage PM$_{2.5}$ predictions by considering these impacts. To estimate missing AOD values, daily gap-filling models with snow/cloud fractions and meteorological covariates were developed using the random forest algorithm. By using these models in New York State, a daily AOD data set with a 1-km resolution was generated with a complete coverage. The “out-of-bag” $R^2$ of the gap-filling models averaged 0.93 with an interquartile range from 0.90 to 0.95. Subsequently, a random forest-based PM$_{2.5}$ prediction model with the gap-filled AOD and covariates was built to predict fully covered PM$_{2.5}$ estimates. A ten-fold cross-validation for the prediction model showed a good performance with an $R^2$ of 0.82. In the gap-filling models, the snow fraction was of higher significance to the snow season compared with the rest of the year. The prediction models fitted with/without the snow fraction also suggested the discernible changes in PM$_{2.5}$ patterns, further confirming the significance of this parameter. Compared with the methods without considering snow and cloud covers, our PM$_{2.5}$ prediction surfaces showed more spatial details and reflected small-scale terrain-driven PM$_{2.5}$ patterns. The proposed methods can be generalized to the areas with extensive snow/cloud covers and large proportions of missing satellite AOD data for predicting PM$_{2.5}$ levels with high resolutions and complete coverage.

*Corresponding author at: Emory University, Rollins School of Public Health, 1518 Clifton Rd. NE, Atlanta, GA 30322, USA. yang.liu@emory.edu (Yang Liu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
1. Introduction

Fine particulate matter (PM$_{2.5}$) may be inhaled and deposited in alveoli, increasing the risk of cardiorespiratory disease (Bose et al. 2015; Burnett et al. 2014; Madrigano et al. 2013; Sorek-Hamer et al. 2016). Though air quality stations have been employed to measure PM$_{2.5}$ and its composition, their spatial coverage is insufficient to cover larger populations to study the impacts of PM$_{2.5}$ exposure on human health. Recently, satellite aerosol optical depth (AOD) with broad spatiotemporal availability has been widely applied in PM$_{2.5}$ exposure modeling (Hu et al. 2017; Kloog et al. 2012; Ma et al. 2016). With a 1-km resolution, Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD has been able to reveal pollution patterns more detailly and link between PM$_{2.5}$ and microenvironment more precisely (Hu et al. 2014; Lyapustin et al. 2011). High-resolution AOD is critical for reflecting PM$_{2.5}$ pollution patterns in the regions dominated by local sources, e.g., New York State, US, in which residential wood combustions and on-road emissions are the primary PM$_{2.5}$ sources (Fine et al. 2002; Reff et al. 2009; Su et al. 2013).

Though satellite AOD can serve as a surrogate of ground PM$_{2.5}$, a non-linear relationship is shown between the two, which varies spatially and temporally (Paciorek and Liu 2009). To reflect this relationship, various statistical models have been developed (Kloog et al. 2014; Liu et al. 2009; Xiao et al. 2017). For instance, Kloog et al. (2014) employed mixed models to estimate high-resolution PM$_{2.5}$ with an out-of-sample $R^2$ (coefficient of determination) of 0.88 in the Northeastern US. Xiao et al. (2017) built two-stage models to produce 1-km PM$_{2.5}$ estimates with complete coverage over the Yangtze River Delta (YRD) in China, which yielded cross-validation (CV) $R^2$s of ~0.8. In these parametric models, however, restrictive assumptions of independence and population distributions are needed, which can be a challenge for a complex mixture, e.g., PM$_{2.5}$. In contrast, non-parametric models (e.g., machine learning models) can capture the non-linear relationships and interactions between the variables with fewer a priori assumptions. Popular machine learning models for PM$_{2.5}$ prediction consist of Artificial Neural Networks (ANN) (Di et al. 2016; Gupta and Christopher 2009; Zou et al. 2015) and Random Forests (RF) (Brokamp et al. 2018; Hu et al. 2017). To be specific, random forests are advantaged in the interpretability of modeling outcomes by their measures of variable importance (Friedman et al. 2001). Hu et al. (2017) developed a random forest model with satellite AOD and covariates, generating 12-km PM$_{2.5}$ predictions over the Contiguous US with a CV $R^2$ of 0.80 and a root-mean-square error (RMSE) of 1.78 μg/m$^3$.

The large proportion of non-random missing data is considered as a critical problem associated with the effective use of satellite AOD, which primarily results from cloud cover and high surface brightness from factors such as snow and ice (Chu et al. 2002; Hsu et al. 2013; Levy et al. 2007). This non-random missingness has significantly affected data availability and generated biases in the stages of data processing (Xiao et al. 2017). Several
strategies have been developed to cope with this data gap problem. van Donkelaar et al. (2011) adopted adjusted cloud filtering standards in the Dark Target algorithm and achieved a 21% increase of available AOD during Moscow wildfire events in 2010. Kloog et al. (2011), based on universal kriging with daily mean PM$_{2.5}$ levels and random slopes, proposed a spatial smoothing model and generated full-coverage PM$_{2.5}$ predictions across New England within the US. Kloog et al. (2012) incorporated inverse probability weighting (IPW) with PM$_{2.5}$ regression to reduce the selection bias caused by AOD missingness. They achieved PM$_{2.5}$ predictions with higher reliability. However, it has been found that AOD and PM$_{2.5}$ levels were changed in the presence of cloud cover due to the shifted AOD/PM$_{2.5}$ physical characteristics under various meteorological conditions (Alam et al. 2014; Belle et al. 2017; Kang et al. 2015; Myhre et al. 2007). For instance, Belle et al. (2017) found that variations in cloud properties (e.g., cloud optical depth and emissivity) were correlated with the shifted levels of PM$_{2.5}$ and its components. Accordingly, the gap-filling process without considering cloud interactions may cause estimation biases. Xiao et al. (2017) showed that the cloud-AOD interactions can be partially explained by incorporating cloud features in AOD gap-filling processes. By using a multiple imputation model coupled with cloud fractions, they found a higher level of gap-filled AOD on cloudy days in the YRD of China. Similar to cloud, snow cover, with a large areal extent in high-latitudes of Northern Hemisphere (Robinson et al. 1993), has also resulted in a large proportion of missing AOD (Lyapustin et al. 2011; Xiao et al. 2016) and the changes of AOD/PM$_{2.5}$ levels (Chen et al. 2012; Emili et al. 2011; Green et al. 2015; Whiteman et al. 2014). Emili et al. (2011) found that due to the contamination of cloud and snow pixels, satellite AOD would be occasionally overestimated. Studies suggested that in mountainous terrains, PM$_{2.5}$ levels were higher on snowy days because of more stagnant atmospheric conditions (Chen et al. 2012; Whiteman et al. 2014). To our knowledge, no study has been conducted to deal with the snow-related AOD missingness in PM$_{2.5}$ modeling. Thus, incorporating snow features in AOD gap-filling processes may help to obtain more reliable PM$_{2.5}$ estimates in the regions with extensive snow cover.

In this case study, we incorporated satellite-retrieved snow and cloud fractions in an AOD gap-filling model to yield fully covered AOD data in New York State where there were extensive snow/cloud covers, especially in winter. Based on the AOD data, we generated high-resolution and fully covered PM$_{2.5}$ predictions. In the meantime, we examined the significance of the snow fraction in the AOD gap-filling and PM$_{2.5}$ prediction processing. We also identified the patterns of local PM$_{2.5}$ pollution in New York State using our high-resolution PM$_{2.5}$ predictions.

2. Data and Methods

2.1. Study Areas

New York State and its surrounding areas were selected as our study areas (Fig. 1). The surrounding areas served as the buffer to minimize the quality degradation on the edge of the areas and to ensure sufficient ground PM$_{2.5}$ stations in the study areas. These areas were suitable for examining the performance of our models since 1) the cloud cover and heavy snowfall led to a considerable amount of missing satellite AOD data, particularly in winter.

Remote Sens Environ. Author manuscript; available in PMC 2020 February 01.
Belle et al. 2017; 2) local PM$_{2.5}$ emission sources (e.g., residential wood combustions and on-road emissions) were dominated in this state so that reliable predictions with a high spatial resolution were needed to reveal the spatial details of PM$_{2.5}$ pollution (Fine et al. 2002; NEI 2014; Su et al. 2013). All analyses were based on the 1-km grid of MAIAC AOD, covering 474,392 grid cells in the entire areas.

2.2. PM$_{2.5}$ Measurements

PM$_{2.5}$ measurements inside the US were provided by the United States Environmental Protection Agency (EPA) (https://www.epa.gov/). The measurements inside Canada were provided by the National Air Pollution Surveillance (NAPS) (http://maps-cartes.ec.gc.ca/). NAPS has three types of samplers (Dichotomous, Partisol and Speciation) for PM$_{2.5}$ measurements. Only Dichotomous and Partisol measurements were employed to ensure the data quality. For each ground station from AQS or NAPS, daily mean PM$_{2.5}$ levels were calculated and matched with AOD and other variables in the MAIAC grid cells. The entire areas covered 137 PM$_{2.5}$ ground stations (with 127 inside the US and 10 inside Canada; Fig. 1).

2.3. MAIAC AOD Data

MAIAC is an advanced AOD data set based on Moderate Resolution Imaging Spectroradiometer (MODIS) Collection 6 Level 1B data. Using a time-series algorithm, the MAIAC data set has increased the accuracies of aerosol detections and optimized the spatial resolution to 1-km (Lyapustin et al. 2011). Since MAIAC is a novel algorithm for MODIS AOD, the quality assessment of this product is insufficient. A current validation over South America suggested that ~66% MAIAC AOD retrievals were within the expected error (±(0.05+0.05×AOD)) (Martins et al. 2017). Terra/Aqua satellites provide two MODIS data sets with different crossing times. In our PM$_{2.5}$ prediction models, Terra (descending node at 10:30 A.M. local time) and Aqua (ascending node at 1:30 P.M. local time) AOD served as two separate predictors to reflect diurnal changes of AOD.

2.4. MODIS Cloud and Snow Fractions

Cloud and snow fractions as the percentages of cloud and snow in a grid cell are critical for the quantitative estimation of cloud- and snow-AOD interactions in the gap-filling processing. Cloud fractions were yielded from MODIS Level-2 Cloud product (MOD06_L2/ MYD06_L2, https://modis.gsfc.nasa.gov/). In this product, the cloud fraction has been retrieved from infrared and visible wavelengths along with other physical and radiative cloud properties (Platnick et al. 2015). A global evaluation by Ackerman et al. (2008) suggested that MODIS cloud detection reached an agreement of ~85% with the ground lidar observations. Snow fractions were yielded from MODIS Snow Cover product (MOD10/ MYD10, https://modis.gsfc.nasa.gov/). In this product, the snow fraction has been yielded from the Normalized Difference Snow Index (NDSI) (Hall and Riggs 2011). MODIS Snow Cover product was found with a high correlation (r = 0.9) with ground-truth observations (Salomonson and Appel 2006). Terra and Aqua data sets were employed respectively and matched with corresponding AOD data sets. Only daytime cloud/snow observations were adopted.
2.5. Meteorological Data

Meteorological variables were obtained from the North American Regional Reanalysis (NARR) (http://www.emc.ncep.noaa.gov/) (Mesinger et al. 2006) and the North American Land Data Assimilation System (NLDAS) (http://ldas.gsfc.nasa.gov/nldas) (Mitchell et al. 2004). NARR had a resolution of 0.25° and a 3-h temporal interval. NLDAS had a resolution of 0.125° and a 1-h temporal interval. Since NLDAS has higher spatiotemporal resolutions, the parameters from this data set were used first. The parameters not provided by NLDAS were extracted from NARR. The meteorological parameters from NLDAS consisted of air temperature, humidity, surface pressure, precipitation, wind speed, potential evaporation, downward shortwave radiation as well as convective available potential energy (CAPE). The parameters from NARR covered planetary boundary layer height (HPBL) and visibility. To match the observation time of MAIAC AOD data, we calculated the daily values of meteorological parameters by averaging the simulations from 9 A.M. to 3 P.M. (local time, i.e., GMT 1400 – 2000 for NLDAS and GMT 1500, 1800, 2100 for NARR).

2.6. Land-Use Variables

The land-use parameters covered (1) the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation at a 1 arc-second (~30 m) resolution (https://asterweb.jpl.nasa.gov/gdem.asp); (2) LandScan ambient population in 2015 at a 900-m resolution (https://web.ornl.gov/sci/landscan/); (3) Normalized Difference Vegetation Index (NDVI) from MODIS vegetation indices at a 500-m resolution (MOD13/MYD13, https://modis.gsfc.nasa.gov/); (4) distances to highways and major roads computed from ESRI StreetMap (Environmental System Research Institute, Inc., Redland, California).

2.7. Data Matching

All data sets with various spatial resolutions were matched and fitted into the MAIAC AOD 1-km grid. For the data sets with spatial resolutions coarser than 1 km (e.g., meteorological fields, cloud and snow fractions, and NDVI), the inverse distance weighting (IDW) interpolation was employed (Bar-tier and Keller 1996). For the data sets with finer resolutions (e.g., ASTER GDEM), the downsampling was conducted, i.e., a process of averaging neighboring pixels into one value. For the LandScan population, each data point was assigned to its nearest MAIAC grid cell. Subsequently, the population in that grid cell was summarized.

2.8. AOD Gap-filling Model

Random Forest (RF) is an “ensemble learning” method yielding several decision trees and aggregating the regressing results from these trees (Breiman 2001; Liaw and Wiener 2002). It adopts two types of bootstrap aggregating strategies (i.e., a bootstrap training sample and a group of randomly selected independent variables) for each decision tree. These strategies allow the random forest to be a robust algorithm against overfitting (Breiman 2001). A random forest model has two major parameters, i.e., the number of decision trees in the forest (ntree) and the number of independent variables in each tree (mtry). It also provides variable importance measures that inform variable weights and contribute to the
interpretation of the model (Liaw and Wiener 2002). Our AOD gap-filling model was based on the random forest algorithm, which is expressed as

$$AOD_{st} = f(\text{cloud fraction}_{st}, \text{snow fraction}_{st}, \text{air temperature}_{st}, \text{specific humidity}_{st}, \text{relative humidity}_{st}, \text{previous day precipitation}_{st}, \text{elevation}_{s}, X_s, Y_s)$$

Equation 1

where $s$ denotes the location of a grid cell, and $t$ is the time of an observation. The dependent variable was the satellite AOD. Independent variables covered cloud and snow fractions, meteorological parameters (i.e., air temperature, specific humidity, relative humidity and precipitation on the previous day), elevation and spatial coordinates. The spatial coordinates were the real distances in kilometers to the central point of the study areas. By comparing the results with different settings, $n_{\text{tree}}$ and $m_{\text{try}}$ were set as 200 and 3 respectively to balance the prediction accuracy and computational efficiency.

The AOD gap-filling model (Equation 1) was fitted daily for both Terra and Aqua AOD data sets. Equation 1 without the snow fraction (a.k.a. cloud-only gap-filling model) was also fitted to examine the impact of this variable on the gap-filled AOD. Since MAIAC AOD had a large proportion of missing data, we employed three rolling-day samples for the middle day’s model. Out-of-bag (OOB) $R^2$ and RMSE were employed to assess the modeling performance. OOB $R^2$ was calculated from the predictions not in the bootstrap sample (a.k.a. “out-of-bag” sample) (Breiman 2001). Since the mechanism of OOB is similar to cross-validation (CV), OOB $R^2$ is nearly equal to CV $R^2$ under large sample size (Friedman et al. 2001). RMSE was calculated by aggregating the errors of OOB predictions. Model-estimated “permutation accuracy importance” (Breiman 2002) was used to suggest the variable importance. This importance measure was estimated in line with the fall of prediction accuracy after randomly permuting the OOB sample of the targeting variable (Liaw and Wiener 2002).

2.9. PM$_{2.5}$ Prediction Model

The PM$_{2.5}$ prediction model also followed the random forest algorithm (Equation 2). The dependent variable was the PM$_{2.5}$ measurements from AQS and NAPS. Independent variables covered gap-filled Terra and Aqua AOD, meteorological parameters (air temperature, dew-point temperature, surface pressure, specific humidity, wind speed, visibility, planetary boundary layer height, potential evaporation, downward shortwave radiation and CAPE), land- use parameters (population, distances to highways and major roads, elevation and NDVI), and dummy variables for months and Julian days. Moreover, a convolutional layer of PM$_{2.5}$ levels was applied, representing the weighted averages of nearby PM$_{2.5}$ levels. This parameter helps to fully exploit the spatial autocorrelation of PM$_{2.5}$ and can significantly increase the accuracy of PM$_{2.5}$ prediction (Di et al. 2016; Hu et al. 2017). By comparing the results with different settings, $n_{\text{tree}}$ and $m_{\text{try}}$ were set as 500 and 7 respectively to achieve the best prediction accuracy.
PM \textsubscript{2.5} \textsubscript{st} = f(\text{Terra AOD}_{\text{st}}, \text{Aqua AOD}_{\text{st}}, \text{air temperature}_{\text{st}}, \text{dew point temperature}_{\text{st}}, \text{surface pressure}_{\text{st}}, \\
\text{specific humidity}_{\text{st}}, \text{wind speed}_{\text{st}}, \text{visibility}_{\text{st}}, \text{HPBL}_{\text{st}}, \text{potential evaporation}_{\text{st}}, \\
\text{download shortwave radiation}_{\text{st}}, \text{CAPE}_{\text{st}}, \text{population}_{\text{st}}, \text{NDVI}_{\text{st}}, \text{highway distance}_{s}, \text{major distance}_{s}, \\
\text{elevation}_{s}, \text{PM2.5 concoloral layer}_{\text{st}}, \text{month}_{t}, \text{day}_{t})

\text{Equation 2}

where \( s \) denotes the location of a grid cell and \( t \) denotes the time of an observation. The variable selection strategy used in Hu et al. (2017) was adopted here, in which the variables with low importance values were discarded (e.g., precipitation and relative humidity in this analysis). 10-fold CVs was employed to assess the model performance, including overall, spatial and temporal CVs (Xiao et al. 2017). The spatial CV generated validation samples in accordance with the locations of the PM\textsubscript{2.5} measurements. The temporal CV generated validation samples in line with the Julian days of the measurements. Besides the original model, Equation 2 without both Terra and Aqua gap-filled AOD parameters (aka. no-AOD prediction model) was fitted to examine the impact of the gap-filled AOD on the PM\textsubscript{2.5} predictions. Equation 2 with AOD gap-filled by cloud fractions (aka. cloud-only prediction model) was also fitted to verify the importance of the snow parameter in the PM\textsubscript{2.5} prediction.

2.10. Comparison with Another Gap-filling Method

To assess our gap-filling model, our full-coverage PM\textsubscript{2.5} predictions were compared to the gap-filled PM\textsubscript{2.5} generated by a previous gap-filling method proposed by Kloog et al. (2011) and Just et al. (2015) (Equation 3).

\[
\sqrt{\text{PredPM}_{\text{st}}} = \alpha_0 + \alpha_1 \sqrt{\text{MeanPM}_{\text{t}}} + s(X,Y)_{s} + \epsilon_{\text{st}} \quad \text{Equation 3}
\]

where \( s \) denotes the location of a grid cell, and \( t \) is the time of an observation. The dependent variable PredPM\textsubscript{st} denotes the existing PM\textsubscript{2.5} predictions at location \( s \) at time \( t \). MeanPM\textsubscript{t} denotes an average PM\textsubscript{2.5} measurement on a day \( t \). \( s(X,Y)_{s} \) is a spline surface of the coordinates of the grid cells. The reason for applying square root was to ensure the positive values of PM\textsubscript{2.5} (Just et al. 2015). We had also tried calculating MeanPM\textsubscript{t} within a 100-km buffer (Kloog et al. 2014). Yet the PM\textsubscript{2.5} distribution showed significant artificial patterns due to the sparse monitoring network so that the buffer was not adopted. The gap-filling model was fitted monthly to better reveal the temporal variations of PM\textsubscript{2.5} (Kloog et al. 2014). A 10-fold CV was employed to examine the modeling performance.
3. Results

3.1. Descriptive Statistics for MAIAC AOD Missingness

The Quality Assessment (QA) flags of MAIAC AOD were employed to infer the rates of missing AOD caused by cloud cover, snow cover and water/ice (Table S1). The average daily missing rates associated with cloud and snow are listed in Table 1 (the contribution of water/ice-related missing AOD was negligible in NYS). Since the missing AOD caused by snow cover was primarily in the first 15 weeks (105 days) of 2015, the average missing rates in this period are separately summarized in the table. This period is also referred to as the “snow season” hereinafter. The overall AOD missing rate was ~90% in 2015, and the cloud-related missing AOD was ~76% of the total data. Though the snow-related missingness only took up ~6% in this year, the percentage increased to ~20% in the snow season. There was a ~10% gap between the overall missing rate and the total rate caused by cloud and snow. This gap resulted from the MAIAC pixels located in the areas outside the sensor scanning range (i.e., no measurement). In general, though the missing AOD was primarily caused by cloud cover, snow-related AOD missingness remained a severe issue in the snow season.

3.2. AOD Gap-filling by Random Forests

By Equation 1, full-coverage AOD was yielded for both Terra and Aqua data sets. Daily RF models had a mean OOB $R^2$ of 0.93 (for both Terra and Aqua data sets) with interquartile ranges (IQR) from 0.91 to 0.95 for Terra and 0.90 to 0.95 for Aqua. The annual AOD distributions and the spatial differences between gap-filled and original AOD are shown in Fig. 2. The overall and gap-filled AOD showed the similar spatial patterns. The peak AOD levels appeared in the Allegheny Plateau of Southern NYS. Some populated areas in NYS (e.g., New York City) also had high-level AOD. For Terra, original and gap-filled AOD had the annual means of 0.11 (IQR: [0.10, 0.12]) and 0.26 (IQR: [0.25, 0.26]), respectively. For Aqua, original and gap-filled AOD had the annual means of 0.09 (IQR: [0.08, 0.10]) and 0.25 (IQR: [0.24, 0.25]), respectively. As Fig. 2 (e) and (f) suggest, the gap-filled AOD was higher than the original AOD throughout the state. This outcome is consistent with that of Belle et al. (2017) and Xiao et al. (2017) who found clear evidences that the hygroscopic growth of aerosol droplets under cloudy and humid weather resulted in higher levels of AOD estimates.

3.3. PM$_{2.5}$ Prediction by Random Forests

3.3.1 Modeling Performance—By Equation 2, 1-km PM$_{2.5}$ predictions with full coverage were yielded. The prediction model showed an overall CV $R^2$ of 0.82 (Fig. 3 (a)) with spatial and temporal CV $R^2$'s of 0.74 and 0.81, respectively. The model had an RMSE of 2.16 $\mu$g/m$^3$, suggesting a good prediction accuracy. Fig. 3 (b) shows the ranking of the variable importance. The PM$_{2.5}$ convolutional layer had the highest importance value, which is consistent with that of Hu et al. (2017) who confirmed an improvement of model accuracy by the PM$_{2.5}$ convolutional layer. Five of the top-seven important variables were land-use terms (e.g., population, distances to major roads, elevation, distances to highways, and NDVI). This outcome could suggest that the PM$_{2.5}$ sources in New York State were primarily local sources, and the transportation of regional pollution was relatively weak, which is consistent with the PM$_{2.5}$ emission inventory of New York State in 2014 (NEI...
2014). Though the gap-filled AOD parameters did not have high importance values, they still contributed to a changed spatial pattern of PM$_{2.5}$ estimates (as shown in Section 3.4.1). We also performed the PM$_{2.5}$ prediction for 2002 – 2012, finding that the modeling performance was stable in terms of CV $R^2$ and variable importance ranking (Table S2).

### 3.3.2. PM$_{2.5}$ Predictions

Fig. 4 (a) shows the annual PM$_{2.5}$ distribution in NYS in 2015 (the PM$_{2.5}$ distributions in different seasons shown in Fig. S1). The PM$_{2.5}$ levels had a mean of 5.30 μg/m$^3$ with an IQR from 4.43 to 6.08 μg/m$^3$. The distribution showed clear patterns of higher PM$_{2.5}$ along with the roads and in the populated areas. These patterns are consistent with the PM$_{2.5}$ emission inventory of NYS (NEI 2014) suggesting that the PM$_{2.5}$ sources were primarily residential wood combustions and on-road emissions. The highest PM$_{2.5}$ levels appeared in the large cities, e.g., New York City, Long Island, and Buffalo. Other major cities, e.g., Albany, Rochester, Yonkers, and Syracuse also had relatively high PM$_{2.5}$ levels. In contrast, the Adirondack Mountains in the Northeastern NYS with the lowest population density showed the lowest PM$_{2.5}$ levels.

Since New York State lies upon the portion of Northeast Appalachians, most of its areas, in particular Upstate New York, are dominated by the mountainous terrain. Our high-resolution PM$_{2.5}$ predictions showed obvious evidence of the correlation between elevation and PM$_{2.5}$ in NYS. This correlation can be partially reflected by the variable importance since the elevation was the 4$^{th}$ important variable in the prediction model. Besides, we found obvious patterns of PM$_{2.5}$ accumulation in the valleys of the Allegheny Plateau in winter. An area in the Allegheny Plateau was selected to highlight the valley accumulation of PM$_{2.5}$. Fig. 5 shows the PM$_{2.5}$ distribution with contours (140-m intervals) in this area in winter (January, February, and December in 2015). Elevated PM$_{2.5}$ levels in the valleys were clearly shown. In these valleys, PM$_{2.5}$ levels were ~1–2 μg/m$^3$ higher than those in the surroundings. These results suggested that our 1-km PM$_{2.5}$ predictions could reflect small-scale PM$_{2.5}$ features driven by local geographical factors, which would be smeared with a coarser resolution.

### 3.3.3. Comparison with Another Gap-filling Approach

To verify the effectiveness of our gap-filling model, we compared our PM$_{2.5}$ predictions to the PM$_{2.5}$ generated using a reported gap-filling method (Equation 3). For a fair comparison, Equation 3 followed a subset of our PM$_{2.5}$ predictions derived from the original MAIAC AOD (i.e., removing the PM$_{2.5}$ predictions generated from the gap-filled AOD). The ten-fold cross-validation $R^2$'s of the monthly models had a mean of 0.69 ranged from 0.60 to 0.78, lower than the OOB $R^2$'s of our gap-filling model. Fig. 4 (b) shows the annual PM$_{2.5}$ distribution generated by Equation 3. Even though the model had full-coverage predictions, the PM$_{2.5}$ were spatially over-smoothed. In contrast, since the major PM$_{2.5}$ sources were local sources in NYS (i.e., residential wood combustions and on-road emissions) (NEI 2014; Su et al. 2013), our PM$_{2.5}$ spatial patterns well captured these well-defined sources.

### 3.4. The Importance of Gap-filled AOD with Snow Cover Parameter

#### 3.4.1. Contribution of Gap-filled AOD to PM$_{2.5}$ Predictions

Since the gap-filled AOD parameters were not among the top-important variables (Fig. 3 (b)), their contribution to our PM$_{2.5}$ predictions was examined to verify their validity in the prediction model. A
prediction model without gap-filled AOD parameters was fitted (a.k.a. no-AOD prediction model), and the analysis of the PM$_{2.5}$ estimates emphasized on the snow season (first 105 days in 2015). Fig. 4 (c) shows the spatial differences between full-model and no-AOD PM$_{2.5}$ in the snow season. Compared with the no-AOD PM$_{2.5}$, the full-model PM$_{2.5}$ had higher levels along with the roads and lower levels in the background areas (e.g., the Adirondack Mountains). Hence, after the addition of the AOD parameters in the prediction model, the pattern of PM$_{2.5}$ sources was intensified, in particular on-road emissions. Quantitatively, the mean absolute difference between two PM$_{2.5}$ data sets was 0.13 μg/m$^3$ with an IQR from 0.06 to 0.18 μg/m$^3$. The maximum absolute difference reached 0.99 μg/m$^3$, suggesting a significant change in terms of a 105-day average. Also, the full-model PM$_{2.5}$ was compared with the PM$_{2.5}$ predictions derived from the original MAIAC AOD (Fig. S2). The significant differences between two data sets suggested the impact of not only the systematic changes of gap-filled AOD values, but also the significant increase of the sample size. These comparisons suggested that the pollution information incorporated in the gap-filled AOD significantly impacted the PM$_{2.5}$ prediction, leading to discernible changes in PM$_{2.5}$ patterns. The reason for the intensified pattern of on-road PM$_{2.5}$ pollution after the addition of the gap-filled AOD needs further studies.

3.4.2. Influences of Snow Cover Parameter on AOD and PM$_{2.5}$—During and after the snow season, the rank change of the snow fraction’s importance in the AOD gap-filling model indicated the importance of this variable in the gap-filling process. Among the 9 variables in the AOD gap-filling model, the average rank of snow fraction’s importance was valued as 5 during the snow season and dropped to 8 after the snow season. In contrast, the rank of cloud fraction’s importance was stable with an average of 2 throughout the year. As expected, this outcome indicated that the snow fraction was more important when it was snowy. Further, we fitted a cloud-only AOD gap-filling model (i.e., the snow fraction was removed) to examine the impact of the snow fraction on the gap-filled AOD (the cross-validation performance listed in Table S3). We found that the mean absolute differences between the full-model AOD and the cloud-only AOD were 0.002 (IQR: [0.001, 0.003]) and 0.001 (IQR: [0.0004, 0.001]) for Terra and Aqua, respectively. During the snow season, the maximum absolute difference was 0.007 in terms of a 105-day average. Moreover, we also examined the impact of the snow fraction on the PM$_{2.5}$ predictions. By using the full-model and cloud-only AOD in the PM$_{2.5}$ prediction, two PM$_{2.5}$ data sets were yielded: 1) full-model PM$_{2.5}$ and 2) cloud-only PM$_{2.5}$. Fig. 4 (d) shows the differences between two PM$_{2.5}$ data sets in the snow season. Compared with the cloud-only PM$_{2.5}$, the full-model PM$_{2.5}$ tended to be higher in the regions with high PM$_{2.5}$ levels (e.g., populated areas and major roads) and lower in backgrounds. Hence, the snow fraction improved the source patterns of PM$_{2.5}$. In the snow season, the mean absolute difference between two PM$_{2.5}$ data sets reached 0.09 μg/m$^3$ (IQR: 0.02 to 0.12 μg/m$^3$), and the maximum absolute difference reached 0.99 μg/m$^3$. The mean difference was at a similar scale as the mean difference caused by the gap-filled AOD, which was 0.13 μg/m$^3$ in the snow season. This suggested that the snow cover might take up the majority of gap-filled AOD’s impact on the PM$_{2.5}$ predictions. In brief, by influencing the gap-filled AOD, the impact of the snow cover was transferred to the PM$_{2.5}$ predictions, resulting in the discernible changes in PM$_{2.5}$ patterns.
The reason for the intensified source patterns after the addition of this parameter needs further studies.

4. Discussion

In this study, fully covered and high-resolution PM$_{2.5}$ levels were estimated based on the gap-filled MAIAC AOD in New York State in 2015. To the best of our knowledge, this is the first study applying both snow and cloud cover parameters in the gap-filling process to increase the spatiotemporal availability of satellite AOD and the accuracy of PM$_{2.5}$ estimation. Despite the ~90% AOD missingness in the region, our gap-filling models still had excellent performance with a mean OOB $R^2$ of 0.93. Xiao et al. (2017) adopted a multiple imputation model with similar predictors (except for the snow fraction) to conduct AOD gap-filling in the YRD of China, in which there was nearly no snow cover, and the AOD missing rate was ~60%. They had a mean modeling $R^2$ of 0.77 with an IQR from 0.71 to 0.82. The improvement in our modeling performance reflected the advantages over the model of Xiao et al. (2017) and the potential of machine learning models dealing with the AOD gap-filling with complex snow/cloud-AOD interactions. Our gap-filled AOD was significantly higher than the original MAIAC AOD by an average of 0.15. This outcome is consistent with the findings of previous studies suggesting that the increased humidity caused by cloud could lead to the aerosol hygroscopic growth (Belle et al. 2017; Kang et al. 2015; Xiao et al. 2017).

Our PM$_{2.5}$ prediction model showed a good performance with a CV $R^2$ of 0.82 and an RMSE of 2.16 $\mu$g/m$^3$, Hu et al. (2017), also applying the random forest algorithm to the PM$_{2.5}$ prediction, had a CV $R^2$ of 0.79 and an RMSE of 2.84 $\mu$g/m$^3$ in the Northeastern US (including New York and New England states). Compared to the model of Hu et al. (2017), our model achieved the major improvements of 1) conducting satellite AOD gap-filling and making full-coverage PM$_{2.5}$ predictions; 2) conducting PM$_{2.5}$ predictions with a higher 1-km spatial resolution; 3) achieving similar modeling performance without the convolutional layers of land-use variables. Kloog et al. (2014) estimated high-resolution PM$_{2.5}$ with an “out-of-sample” $R^2$ of 0.88 and an RMSE of 2.33 $\mu$g/m$^3$ in the Northeastern US using a linear mixed effects model. However, their PM$_{2.5}$ gap-filling model, virtually a generalized additive model (GAM) with a spline surface (Equation 3), tended to over-smooth the PM$_{2.5}$ spatial details. In contrast, our PM$_{2.5}$ predictions showed strong local sources in the populated areas and major roads. 2014 National Emissions Inventory (NEI 2014) suggested that the largest PM$_{2.5}$ emission source in New York State was the residential wood combustion, resulting in 17,916 tons of PM$_{2.5}$ emissions. This emission source was primarily gained around the large and densely populated cities (e.g., New York metropolitan area, Buffalo, Rochester, Syracuse and Albany) (Reff et al. 2009). Thus, the correlation between high PM$_{2.5}$ levels and high populations in our areas was reasonable. Besides, the emission inventory (NEI 2014) suggested that among the top-twelve PM$_{2.5}$ emission sources in New York State (out of 60 Emissions Inventory System (EIS) emission sectors), at least four of them had the direct correlation with on-road emissions. These sources (paved road dust, on-road diesel heavy-duty vehicles, unpaved road dust, and on-road non-diesel light-duty vehicles) led to 19,753 tons of PM$_{2.5}$ emissions in 2014. These high-level PM$_{2.5}$ emissions could interpret the strong PM$_{2.5}$ signals along with the roads in our areas. Due to
the 1-km resolution, our predictions could reflect small-scale PM$_{2.5}$ features driven by local topographic factors. In some mountainous areas of Upstate New York, 1–2 μg/m$^3$ higher PM$_{2.5}$ accumulated in valleys in winter was found. This phenomenon is consistent with the findings of previous studies suggesting that the stagnant weather conditions in valleys caused by strong temperature inversions could lead to the increase in PM$_{2.5}$ levels (Gillies et al. 2010; Silcox et al. 2012; Wang et al. 2012).

This study considered snow cover in AOD gap-filling process by introducing the snow fraction in the gap-filling model. The significance of the snow fraction was partially reflected by its importance ranking in the gap-filling models. In the snow season (first 15 weeks of 2015), when the heaviest snowfalls appeared, the snow fraction would be the 5$^{th}$/6$^{th}$ important variable, compared with the rest of the year when its importance only ranked 8$^{th}$/9$^{th}$. To further examine the impact of the snow fraction on the PM$_{2.5}$ predictions, two prediction models without the gap-filled AOD or snow fraction were built. In the snow season, the gap-filled AOD alone caused the absolute changes of PM$_{2.5}$ by an average of 0.13 μg/m$^3$, and the snow fraction resulted in an average absolute change of 0.09 μg/m$^3$. Accordingly, the impact of the snow fraction might be responsible for the primary influence of gap-filled AOD on the PM$_{2.5}$ predictions in the snow season. For the spatial pattern, the snow fraction led to the increase of PM$_{2.5}$ estimates in the areas with large PM$_{2.5}$ sources (e.g., populated areas and major roads). Previous studies found that the elevated PM$_{2.5}$ levels appeared to occur on snowy days due to more stagnant atmospheric conditions (Chen et al. 2012; Whiteman et al. 2014). The impact of snow on PM$_{2.5}$ patterns in this analysis could partially reflect this phenomenon. On the whole, by affecting the gap-filled AOD, the impact of the snow fraction was transferred to the PM$_{2.5}$ predictions, leading to the discernible changes in PM$_{2.5}$ patterns. These analyses also suggested the advantages of estimating missing satellite AOD data for PM$_{2.5}$ prediction, instead of directly estimating the missing PM$_{2.5}$ levels. The former approach can, to a greater extent, draw upon the pollution characteristics incorporated in the satellite AOD and the additional information provided by AOD-related meteorological features. Though this study was limited to New York State, the methodology relative to the AOD gap-filling and PM$_{2.5}$ prediction is generalizable to other areas with extensive snow/cloud covers and large proportions of missing satellite AOD data.

The major limitation of this study is that the physical characteristics of snow and cloud are insufficiently considered. The parameters used in this study only reflected the coverage of snow and cloud. Their different physical features, however, may cause the changes in the AOD and PM$_{2.5}$ levels (Belle et al. 2017). Accordingly, more snow/cloud characteristics (e.g., cloud optical depth, cloud emissivity, surface albedo, etc.) can be applied to better interpret the interactions between snow/cloud and AOD/PM$_{2.5}$. With an increased number of snow/cloud parameters, a more suitable strategy to incorporate them, instead of additively applying them to the model, also deserves to be considered. Furthermore, the reason for snow/cloud-related AOD/PM$_{2.5}$ pattern changes needs further studies.

5. Conclusion

In this study, an AOD gap-filling model and a PM$_{2.5}$ prediction model based on the random forest algorithm were developed to estimate fully covered and high-resolution ground PM$_{2.5}$
in New York State in 2015. By introducing the MODIS snow/cloud fractions into the gap-filling process, a 100% gap-filled AOD data set was produced with an excellent modeling performance. The 1-km PM$_{2.5}$ predictions derived from the gap-filled AOD could reflect the detailed emission patterns and small-scale terrain-driven features. It is the first attempt where both snow and cloud parameters are introduced into the AOD gap-filling process. Though we only applied fraction measures of snow and cloud, the importance of these parameters was still reflected, and the discernible interactions between snow/cloud and AOD/PM$_{2.5}$ were observed. It is necessary for future applications to adopt more physical characteristics of snow and cloud and to explore more suitable strategies to introduce these parameters into the gap-filling process. The methodology of this study can be generalized to other areas with extensive snow/cloud covers and large proportions of missing satellite AOD data to estimate PM$_{2.5}$ exposures that previously could not be obtained. The improved PM$_{2.5}$ exposures with an increased sample size and good data quality are expected to contribute to downward epidemiological studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work of J. Bi, J. Belle, and Y. Liu is partially supported by the NASA Applied Sciences Program (Grant # NNX14AG01G and NNX16AQ28G, PI: Liu). This work is also supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number P30ES019776. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Breiman L (2002). Manual on setting up, using, and understanding random forests v3. 1 Statistics Department University of California Berkeley, CA, USA, 1

Remote Sens Environ. Author manuscript; available in PMC 2020 February 01.
burden of disease attributable to ambient fine particulate matter exposure. Environ Health Perspect, 122, 397–403 [PubMed: 24518036]


Silcox GD, Kelly KE, Crosman ET, Whitteman CD, & Allen BL (2012). Wintertime PM2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley. Atmospheric Environment, 46, 17–24


Whiteman CD, Hoch SW, Horel JD, & Charland A (2014). Relationship between particulate air pollution and meteorological variables in Utah’s Salt Lake Valley. Atmospheric Environment, 94, 742–753


Highlights:

• Snow and cloud features were incorporated into satellite AOD gap-filling process
• Reliable PM2.5 predictions with full coverage were generated at a 1-km resolution
• Snow and cloud covers led to discernible changes in PM2.5 pollution patterns
• Small-scale terrain-driven PM2.5 pollution patterns were found in valleys in winter
Fig. 1.
Study areas. Latitude: [40.1°N, 45.6°N]; Longitude: [80.5°W, 71°W]. The area outside New York State served as the buffer. Red diamonds are EPA AQS stations. Blue triangles are NAPS stations.
Fig. 2.
AOD spatial distributions in 2015: (a) annual distribution of Aqua AOD (original and gap-filled AOD); (b) annual distribution of gap-filled Aqua AOD; (c) annual distribution of Terra AOD (original and gap-filled AOD); (d) annual distribution of gap-filled Terra AOD; (e) differences between gap-filled and original Aqua AOD (gap-filled minus original AOD); (f) differences between gap-filled and original Terra AOD (gap-filled minus original AOD).
Fig. 3.
Random forest modeling performance: (a) 10-fold cross-validation (CV) scatters with an $R^2$ of 0.82 and an RMSE of 2.16 μg/m$^3$; (b) variable importance ranking. The PM$_{2.5}$ convolutional layer had the highest value and five of the top-seven important variables were land-use terms.
Fig. 4.
PM$_{2.5}$ spatial distributions in 2015: (a) annual distribution of PM$_{2.5}$ with a 1-km resolution by Equation 2; (b) gap-filled PM$_{2.5}$ by Equation 3; (c) differences between full-model and no-AOD PM$_{2.5}$ in the snow season (full-model minus no-AOD PM$_{2.5}$); (d) differences between full-model and cloud-only PM$_{2.5}$ in the snow season (full-model minus cloud-only PM$_{2.5}$).
Fig. 5.
PM$_{2.5}$ accumulation effect in the valleys of Upstate New York in winter. The areas are on the border of Chenango County and Otsego County (Latitude: [42.375°N, 42.75°N]; Longitude: [75.5°W, 75.15°W]). The widths of the valleys are ~3 km.
Table 1:
Average daily missing rates of MAIAC AOD caused by cloud and snow covers in NYS in 2015. The first 15 weeks of 2015 are defined as the “snow season”. The ~10% gap between the overall missing rate and the total rate caused by cloud and snow resulted from the MAIAC pixels outside the sensor scanning range (i.e., no measurement). The contribution of water/ice to the missing AOD was negligible in NYS.

<table>
<thead>
<tr>
<th>AOD</th>
<th>Missing Type</th>
<th>Mean</th>
<th>Median</th>
<th>25th - 75th Quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqua</td>
<td>Overall</td>
<td>90.27%</td>
<td>96.54%</td>
<td>87.62% - 99.84%</td>
</tr>
<tr>
<td></td>
<td>Cloud-related</td>
<td>75.58%</td>
<td>79.47%</td>
<td>64.74% - 91.48%</td>
</tr>
<tr>
<td></td>
<td>Snow-related</td>
<td>6.14%</td>
<td>0.0005%</td>
<td>0% - 2.48%</td>
</tr>
<tr>
<td></td>
<td>Snow-related (first 15 weeks)</td>
<td>21.15%</td>
<td>14.03%</td>
<td>5.23% - 31.18%</td>
</tr>
<tr>
<td>Terra</td>
<td>Overall</td>
<td>89.50%</td>
<td>96.54%</td>
<td>84.88% - 99.75%</td>
</tr>
<tr>
<td></td>
<td>Cloud-related</td>
<td>76.66%</td>
<td>81.32%</td>
<td>66.29% - 92.12%</td>
</tr>
<tr>
<td></td>
<td>Snow-related</td>
<td>5.55%</td>
<td>0.001%</td>
<td>0% - 2.88%</td>
</tr>
<tr>
<td></td>
<td>Snow-related (first 15 weeks)</td>
<td>19.03%</td>
<td>10.78%</td>
<td>5.00% - 29.84%</td>
</tr>
</tbody>
</table>