Atmos. Chem. Phys., 19, 1097–1113, 2019
https://doi.org/10.5194/acp-19-1097-2019


Atmospheric Chemistry and Physics
Open Access
EGU

# Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea

**Seohui Park[1], Minso Shin[1], Jungho Im[1], Chang-Keun Song[1], Myungje Choi[2,3], Jhoon Kim[2], Seungun Lee[4], Rokjin Park[4], Jiyoung Kim[5], Dong-Won Lee[6], and Sang-Kyun Kim[6]**

[1]School of Urban & Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, 44919, Republic of Korea
[2]Department of Atmospheric Sciences, Yonsei University, Seoul, 03722, Republic of Korea
[3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
[4]School of Earth and Environmental Sciences, Seoul National University, Seoul, 08826, Republic of Korea
[5]Global Environment Research Division, Climate and Air Quality Research Department, National Institute of Environmental Research, Incheon, 22689, Republic of Korea
[6]Environmental Satellite Centre, Climate and Air Quality Research Department, National Institute of Environmental Research, Incheon, 22689, Republic of Korea

**Correspondence:** Jungho Im (ersgis@unist.ac.kr)

Received: 28 June 2018 – Discussion started: 13 September 2018
Revised: 15 January 2019 – Accepted: 17 January 2019 – Published: 28 January 2019

**Abstract.** Long-term exposure to particulate matter (PM) with aerodynamic diameters < 10 ($PM_{10}$) and 2.5 µm ($PM_{2.5}$) has negative effects on human health. Although station-based PM monitoring has been conducted around the world, it is still challenging to provide spatially continuous PM information for vast areas at high spatial resolution. Satellite-derived aerosol information such as aerosol optical depth (AOD) has been frequently used to investigate ground-level PM concentrations. In this study, we combined multiple satellite-derived products including AOD with model-based meteorological parameters (i.e., dew-point temperature, wind speed, surface pressure, planetary boundary layer height, and relative humidity) and emission parameters (i.e., NO, $NH_3$, $SO_2$, primary organic aerosol (POA), and HCHO) to estimate surface PM concentrations over South Korea. Random forest (RF) machine learning was used to estimate both $PM_{10}$ and $PM_{2.5}$ concentrations with a total of 32 parameters for 2015–2016. The results show that the RF-based models produced good performance resulting in $R^2$ values of 0.78 and 0.73 and root mean square errors (RMSEs) of 17.08 and 8.25 µg $m^{-3}$ for $PM_{10}$ and $PM_{2.5}$, respectively. In particular, the proposed models successfully estimated high PM concentrations. AOD was identified as the most significant for estimating ground-level PM concentrations, followed by wind speed, solar radiation, and dew-point temperature. The use of aerosol information derived from a geostationary satellite sensor (i.e., Geostationary Ocean Color Imager, GOCI) resulted in slightly higher accuracy for estimating PM concentrations than that from a polar-orbiting sensor system (i.e., the Moderate Resolution Imaging Spectroradiometer, MODIS). The proposed RF models yielded better performance than the process-based approaches, particularly in improving on the underestimation of the process-based models (i.e., GEOS-Chem and the Community Multiscale Air Quality Modeling System, CMAQ).

## 1 Introduction

Epidemiological studies have consistently shown that negative human health effects including premature mortality can be caused by long-term exposure to atmospheric aerosols and particles, especially $PM_{10}$ and $PM_{2.5}$ (particulate matter (PM) with an aerodynamic diameter of less than 10 and 2.5 µm, respectively) (Pope III et al., 2009; Bartell et al.,

2013; Jerrett et al., 2017). Consequently, the monitoring and assessment of exposure to $PM_{10}$ and $PM_{2.5}$ are crucial for effective management of public health risks. In recent decades, East Asia has been significantly industrialized and urbanized through its rapid economic growth. The industrialization and urbanization have resulted in adverse effects on air quality not only in this region but also in neighboring countries (Koo et al., 2012).

The Public Health and Environment Research Institute in South Korea has been monitoring $PM_{10}$ and $PM_{2.5}$ concentrations at numerous sites all over its jurisdiction. Even though the distribution of the monitoring sites is relatively dense, there is a limitation in providing spatially continuous PM concentrations that focus on major urban areas. For example, Zang et al. (2017) studied the effect of a temperature inversion layer on the relationship between aerosol optical depth (AOD) and $PM_{2.5}$. The aerosol robotic network (AERONET) AOD and radiosonde data were used to estimate ground $PM_{2.5}$ concentrations through an optimized subset regression model. They found the temperature inversion layer to be a key factor in enhancing the accuracy of a ground-level $PM_{2.5}$ estimation model with a coefficient of determination ($R^2$) of 0.63 and a root mean square error (RMSE) of 35.45 $\mu g\,m^{-3}$ (Zang et al., 2017). Their study suggested an inversion model to estimate $PM_{2.5}$ but showed a limitation in that the model can only be used in areas near ground stations, which are required by the model to derive its parameters. Ground-based data typically have uncertainty for spatial distribution of PM concentrations as they are point-based measurements requiring spatial interpolation. Satellite-based PM monitoring has the potential to provide information on air quality over vast areas at high spatial resolution. Many studies have examined the use of satellite-based products to estimate surface PM concentrations (Liu et al., 2005; Gupta and Christopher, 2009a, b; Van Donkelaar et al., 2010, 2015; Chudnovsky et al., 2014; Li et al., 2015; Xu et al., 2015a; You et al., 2015; Wu et al., 2016). AOD is the most widely used parameter that can be derived from satellite remote sensing to estimate ground-level PM concentrations. It represents the amount of light attenuation caused by atmospheric aerosol scattering and absorption in the vertical column.

Early studies generally adopted simple linear regression to investigate the relationship between total column AOD and surface PM concentrations (Liu et al., 2005, 2007). Liu et al. (2005) estimated ground-level $PM_{2.5}$ concentrations over the eastern United States using Multiangle Imaging Spectroradiometer (MISR)-derived AOD, planetary boundary layer height (PBLH) and relative humidity (RH) from the Goddard Earth Observing System (GEOS-3). Their results yielded an $R^2$ of 0.48 and an RMSE of 13.8 $\mu g\,m^{-3}$ when the estimated $PM_{2.5}$ concentrations were compared to in situ measurements. Chemical transport models (CTMs) have also been combined with satellite observations to estimate ground-level PM concentrations. To estimate global 6-year (2001–2006) averaged $PM_{2.5}$ concentrations, Van Donkelaar et al. (2010) combined Moderate Resolution Imaging Spectroradiometer (MODIS) and MISR-derived AODs, and multiplied them by the ratio between $PM_{2.5}$ and AOD simulated by the GEOS-Chem model (i.e., CTM). Their results showed a strong spatial agreement with in situ $PM_{2.5}$ concentrations in North America (slope = 1.07; $R^2 = 0.59$).

More recent studies explored advanced statistical and machine learning approaches to improve the prediction of ground-level PM concentrations by deploying mixed-effect models, geographically weighted regression (GWR), support vector machines (SVMs), or artificial neural networks (ANNs) (Gupta and Christopher, 2009b; You et al., 2015; Li et al., 2017a; Chen et al., 2018). Machine learning approaches have been widely used in various remote-sensing studies thanks to their flexibility with classification and regression (Im et al., 2009; Lu et al., 2011a, Liu et al., 2015; Ke et al., 2016; Pham et al., 2017; Forkuor et al., 2018). In particular, random forest (RF) has proved to be useful for remote-sensing-based regression tasks (Yoo et al., 2012, 2018; Jang et al., 2017; Richardson et al., 2017). To estimate daily $PM_{2.5}$ concentrations over the United States, Hu et al. (2017b) incorporated MODIS AOD, simulated GEOS-Chem AOD, meteorological data, and land use information in an RF model. The developed RF model produced an $R^2$ of 0.8 and an RMSE of 2.83 $\mu g\,m^{-3}$ from 10-fold cross-validation.

Most previous studies have mainly used AOD produced from polar orbiting satellite sensor systems such as MODIS and MISR. They provide AOD worldwide but only make it available once a day because of the revisit time. A major problem with daily AOD is cloud contamination. Therefore, it is difficult to obtain spatially continuous AOD over cloudy regions such as East Asia during the summer monsoon. AOD produced from geostationary satellite sensor systems may be a better option for estimating ground-level PM concentrations due to it having a higher temporal resolution than polar orbiting sensor systems. The Geostationary Ocean Color Imager (GOCI) is the world's first geostationary ocean color satellite sensor that provides multispectral aerosol data in northeast Asia (included eastern China, the Korea peninsula, and Japan) (Park et al., 2014; Xu et al., 2015a). GOCI provides hourly data at 500 m resolution eight times a day from 09:00 to 16:00 Korean Standard Time (KST). Xu et al. (2015a) examined $PM_{2.5}$ concentrations in eastern China using GOCI-derived AOD, coupled with GEOS-Chem simulation data, resulting in a strong correlation ($R^2 = 0.66$) with in situ measurements in terms of annual mean concentrations.

In addition, recent studies have used PBLH, RH, wind speed, and other meteorological variables and land use information because these factors are related to PM concentrations and thus can be used to improve estimation models (Gupta and Christopher, 2009a; Liu et al., 2009; Wu et al., 2012, 2016; Chudnovsky et al., 2014; You et al., 2015; Li et al., 2017b; Yeganeh et al., 2017). In this study, we adopted

the machine learning approach, RF, to develop models estimating ground-level $PM_{10}$ and $PM_{2.5}$ concentrations using satellite-derived products, numerical and emission model output, and ancillary spatial data over South Korea. Aerosol products retrieved from GOCI including AOD were used as key input variables. The objectives of this study are to (1) estimate ground-level $PM_{10}$ and $PM_{2.5}$ concentrations based on GOCI aerosol products and meteorological and emission model output data using RF; (2) validate the estimated PM concentrations using in situ observation data; (3) compare the results to those when MODIS aerosol products were used instead of GOCI products; and (4) evaluate the proposed remote-sensing-based models in comparison with the results from physical models such as GEOS-Chem and the Community Multiscale Air Quality Modeling System (CMAQ).

## 2 Study area and data

### 2.1 Study area

The study area was South Korea (33–39° N, 124–131.5° E), located in northeast Asia, a region known to have relatively poor air quality. Our study area is located in the midlatitude region where the prevailing westerlies carry particulates from the two most rapidly developing countries in Asia (i.e., China and India). The annual mean temperature of South Korea ranges from 10 to 15 °C, and the annual precipitation ranges from 1000 to 1900 mm. More than half of the precipitation occurs in summer during the Asian monsoon. Wind direction is seasonal, with northwesterly winds prevailing in winter and southwesterly winds in summer.

### 2.2 Data

Data used in this study are ground observations as the target variable and remote-sensing data, model-based data, and other ancillary spatial data as explanatory variables. We selected the explanatory variables considering the recent literature that estimated ground PM concentrations (He and Huang, 2018; Chen et al., 2018; Brokamp et al., 2018), which are explained in the following sections.

#### 2.2.1 Observation data

PM observation data (i.e., $PM_{10}$ and $PM_{2.5}$) in South Korea were obtained from the AirKorea website (https://www.airkorea.or.kr/, last access: 24 January 2019) for the period from 2015 to 2016. A total of 325 stations are distributed throughout the country with a concentration in metropolitan areas such as the Seoul Metropolitan Area (SMA) (Fig. 1). Hourly concentrations of air pollutants such as $PM_{10}$ and $PM_{2.5}$ are provided as real-time data. PMs at stations are measured based on a beta attenuation monitoring (BAM) technique, which is widely used for automatic air monitoring (Zhan et al., 2016, 2017). The measurement results are expressed as mass concentration per unit volume (i.e., $\mu g\,m^{-3}$) converted to room temperature (20 °C, 1 atm). Currently, $PM_{10}$ data are provided at 316 stations while $PM_{2.5}$ are measured at 194 stations.

#### 2.2.2 Remote-sensing data

Various remote-sensing data were used in this study such as GOCI aerosol products, the MODIS Normalized Difference Vegetation Index (NDVI), a land cover product, Global Precipitation Measurement (GPM) 30 min precipitation data, and the Shuttle Radar Topography Mission (SRTM) elevation data. GOCI is a geostationary satellite imaging sensor onboard the Communication, Ocean, and Meteorological Satellite (COMS), which was launched in June 2010. It covers 2500 km × 2500 km over the East Asia region, and eight images collected at six visible and two near-infrared (NIR) bands per day are provided hourly from 09:00 to 16:00 in local time (KST). GOCI aerosol products are derived by the GOCI Yonsei aerosol retrieval (YAER) version 2 algorithm (Choi et al., 2018). Four types of products were used in this study: AOD at 550 nm, fine-mode fraction (FMF) at 550m, single-scattering albedo (SSA) at 440 nm, and the Ångström exponent (AE) at 440 and 870 nm with a 6 km × 6 km spatial resolution (Table 1).

The MODIS satellite instrument, onboard the Terra and Aqua satellites, acquires data in 36 spectral bands ranging from 0.4 to 1.4 µm in wavelength. The 16-day NDVI with 1 km resolution (MYD13A2; Solano et al., 2010), aerosol 5 min L2 swath data with 3km resolution (MYD04_3K; Levy et al., 2013) products from 2015 to 2016, and the yearly land cover type product with 500 m resolution (MCD12Q1; Friedl et al., 2010) in 2013 were obtained from Earthdata (https://search.earthdata.nasa.gov/, last access: 24 January 2019). Urban area ratios were calculated using land cover data based on the 13 × 13 neighborhood pixels, which were similar to the spatial resolution of GOCI AOD products. The MODIS aerosol product was used for comparison with GOCI AOD data.

The GPM (Huffman et al., 2015) developed by the National Aeronautics and Space Administration (NASA) and the Japanese Aerospace Exploration Agency (JAXA), was launched in February 2014 to provide observations of rain and snow worldwide. Half-hourly precipitation data with 0.1° resolution (3IMERGHH) were obtained from Goddard Earth Science Data and Information Service Centre (GES DISC; https://mirador.gsfc.nasa.gov/, last access: 24 January 2019). Half-hourly precipitation data were provided as precipitation rates with $mm\,h^{-1}$ and used to calculate 24 h accumulated precipitation data for every hour.

The SRTM (Farr et al., 2007) was launched as a payload on the STS-99 mission of the Space Shuttle *Endeavour* to generate a global digital elevation model (DEM) of the Earth. SRTM DEM data were acquired using the radar interferometry based on the C-band Spaceborne Imaging Radar (SIR-

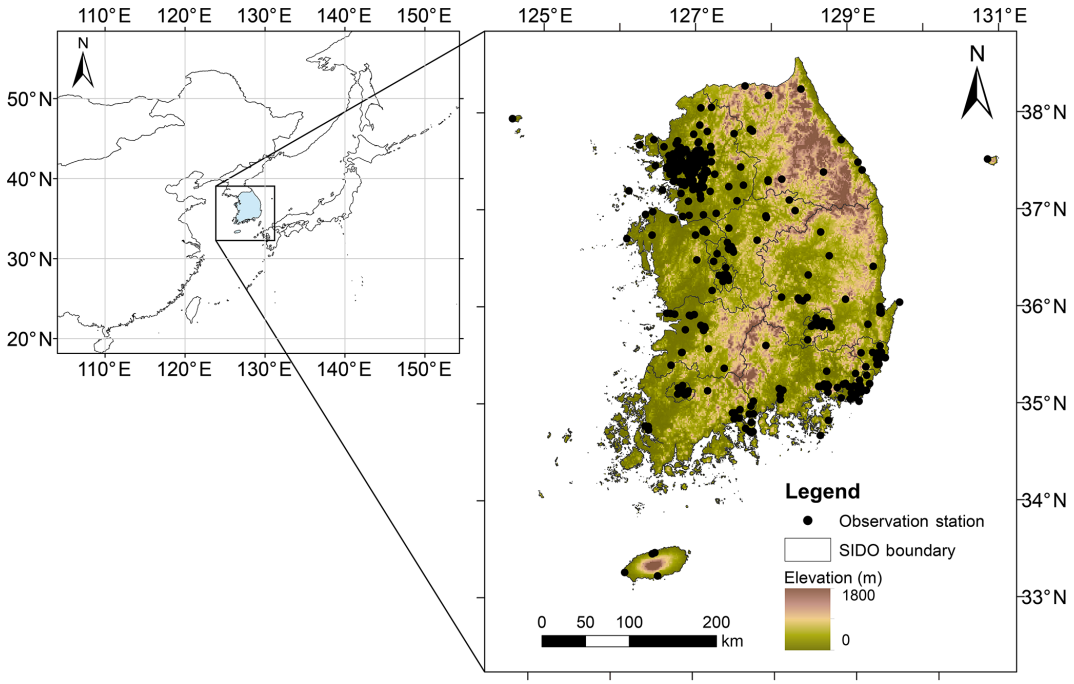


**Figure 1.** Study area with particulate matter (PM) monitoring station sites in South Korea. Elevation is used as a background image.

C) and the X-band Synthetic Aperture Radar (X-SAR) hardware. The elevation data were provided at 1 (about 30 m) and 3 arcsec (about 90 m) spatial resolution for global coverage from the US Geological Survey (USGS) EarthExplorer website (https://earthexplorer.usgs.gov/, last access: 24 January 2019). In this study, 3 arcsec data were used and resampled to the same resolution as the MODIS data with 1 km spatial resolution (Table 1).

### 2.2.3 Model-based data

Along with satellite-based data, the outputs from three models were combined. The three models were the Regional Data Assimilation and Prediction System (RDAPS), the Sparse Matrix Operator Kernel Emissions (SMOKE), and the Breathing Earth System Simulator (BESS). The RDAPS (Davies et al., 2005) is one of the numerical weather forecast models used by the Korea Meteorological Administration, which is based on the Unified Model (UM) developed by the United Kingdom Met Office. The spatial domain of the RDAPS is 77.38–176.56° E and 9.59–61.27° N. The RDAPS takes the information of initial and boundary conditions from the UM Global Data Assimilation and Prediction System (GDAPS) with a spatial resolution of 25 km × 25 km. The analysis–forecast products with about a 100 variables are generated with 12 km spatial resolution and 70 vertical layers. They are provided four times a day (03:00, 09:00, 15:00, 21:00 KST) for 87 h forecasts with 3 h time steps. A total of seven variables in UM RDAPS analysis data (i.e., temperature, dew-point temperature, RH, maximum wind speed, visibility at the height above the ground, PBLH, and surface pressure) were used as meteorological input variables in this study. These meteorological variables are commonly used to estimate ground-level PM concentrations (Lv et al., 2017; He and Huang, 2018).

SMOKE (Baek et al., 2009) is based on emission inventories generally provided as an annual total emission amount for each emission source. Hourly emission data with 9 km spatial resolution were obtained from the National Institute of Environmental Research (NIER). Among the 47 chemical composition parameters in SMOKE outputs, 14 PM-related emission data parameters (i.e., ISOPRENE, TRP1, $CH_4$, NO, $NO_2$, $NH_3$, HCOOH, HCHO, CO, $SO_2$, primary organic aerosol (POA), $PNO_3$, $PSO_4$, and other primary $PM_{2.5}$ (PMFINE) were used in this study. The selected parameters are mostly those defined by Aerosol Emission 5 (AE5) as major precursors forming the PM (Xu et al., 2015b; van Zelm et al., 2016; Gao et al., 2016).

BESS (Ryu et al., 2018) is the MODIS-based model that couples atmosphere and canopy radiative transfers, photosynthesis, transpiration, and energy balance. It includes an atmospheric radiative transfer model and an ANN approach with MODIS atmospheric products. Daily BESS shortwave radiation products with 5 km spatial resolution were obtained from the Environmental Ecology Lab at Seoul National University (http://environment.snu.ac.kr/bess_rad/, last access: 24 January 2019).

**Table 1.** Remote-sensing data used to develop models estimating ground-level particulate matter concentrations in this study.

| Product | Spatial resolution | Temporal resolution | Variables | Description |
|---|---|---|---|---|
| GOCI AOD_550nm | 6 km | 8 day$^{-1}$ | Aerosol optical depth (AOD) | The measure of the extinction of the solar radiation by aerosols (e.g., dust, haze, and sea salt) |
| GOCI FMF_550nm | 6 km | 8 day$^{-1}$ | Fine-mode fraction (FMF) | The ratio of small-size aerosols (radii between 0.1 and 0.25) to the total aerosols |
| GOCI SSA_440nm | 6 km | 8 day$^{-1}$ | Single-scattering Albedo (SSA) | The measure of the amount of aerosol light extinction due to scattering |
| GOCI AE_440_870nm | 6 km | 8 day$^{-1}$ | Ångström exponent (AE) | The exponent related to particle size (the smaller the particles, the bigger the Ångström exponent) |
| MODIS MYD13A2 | 1 km | 16 days | Normalized Difference Vegetation Index (NDVI) | The indicator denoting vegetation quantification |
| MODIS MCD12Q1 | 500 m | yearly | Land cover type (urban area ratio) | The ratio of urban area to 6 km × 6 km neighborhood of each pixel |
| GPM 3IMERGHH | 0.1° | 30 min | Precipitation | The 24 h accumulated precipitation produced using 30 min 3MERGHH precipitation data from GPM |
| SRTM void filled | 90 m | – | Digital elevation model (DEM) | The 2-D representation of topographic surface |

#### 2.2.4 Other input variables

Population density by region (obtained from the Statistical Geographic Information Service (SGIS; https://sgis.kostat.go.kr/, last access: 24 January 2019)) and day of year (DOY) were used as additional input variables together with remote-sensing- and model-based meteorological and emission variables. Population density was calculated for each administrative division, in which a unit is the number of people per square kilometer, and then converted to raster with a 1 km grid. In this study, DOY was converted to values ranging from −1 to 1 with a 1-year period using a sine function considering seasonality (i.e., setting the middle of summer as 1 and the middle of winter as −1; Stolwijk et al., 1999). Road network data were not used in this study, as the use of the road data often yielded inaccurate results over nonurban areas in our preliminary analyses.

#### 2.2.5 Data preprocessing

A total of 32 input variables from satellite- and model-based data were used for the estimation of ground-level PM concentrations in the RF machine learning. All data collected at 13:00 KST were used to develop PM estimation models to match the acquisition time of MODIS Aqua aerosol products over the study area. The observed PM concentrations (i.e., target variables) were log-transformed because the concentration range is large and has a positively skewed distribution. To ensure the reliability of GOCI-derived aerosol products, the four rule-based filters used in Choi (2017) were applied: buddy check, local variance check, sub-pixel cloud fraction check, and diurnal variation check. The same NDVI values during the interval of the MODIS 16-day NDVI were used in the models. GPM precipitation data were converted into 24 h accumulated precipitation data using 48 half-hourly data prior to the target time (i.e., hourly). UM RDAPS reanalysis data were linearly interpolated using analysis fields at 09:00 and 15:00 KST. DEM, urban area ratio, and population density data were used as constant variables during the study period. Input data with different spatial resolutions were resampled to a 1 km MODIS grid using bilinear interpolation. A total of 32 input variables and their abbreviations are summarized in Table 2.

## 3 Methodology

The process flow diagram for the estimation of ground-level PM concentrations is shown in Fig. 2. The constructed data were divided into two groups by date: 80 % of the data were used for model development and the remaining 20 % were used for hindcast validation considering data distribution by PM concentration levels. The data for model development were again randomly divided into training (80 %) and test (20 %) datasets. Since PM reference data had a skewed distribution (i.e., a number of low-concentration samples and a few high-concentration samples), oversampling and subsampling approaches were conducted only for the training dataset

**Table 2.** List of input variables (and their abbreviations) used to estimate ground-level particulate matter concentrations.

| Data | Variables | Abbreviations |
|---|---|---|
| Satellite-based remote-sensing data | Aerosol optical depth | AOD |
| | Fine-mode fraction | FMF |
| | Single-scattering albedo | SSA |
| | Ångström exponent | AE |
| | Normalized Difference Vegetation Index | NDVI |
| | Urban area ratio | Urban_ratio |
| | 24 h accumulated precipitation | Precip |
| | Digital elevation model | DEM |
| Model-based meteorological data | Temperature at the height above ground | Temp |
| | Dew-point temperature at the height above ground | Dew |
| | Relative humidity at the height above ground | RH |
| | Pressure surface | P_srf |
| | Three-hour maximum wind speed at the height above ground | MaxWS |
| | Planetary boundary layer height | PBLH |
| | Visibility at the height above ground | Visibility |
| | Solar radiation | RSDN |
| Model-based emission data | Isoprene ($C_5H_8$) | ISOPRENE |
| | Monoterpene ($C_{10}H_{16}$) | TRP1 |
| | Methane ($CH_4$) | $CH_4$ |
| | Nitric oxide (NO) | NO |
| | Nitrogen dioxide ($NO_2$) | $NO_2$ |
| | Ammonia ($NH_3$) | $NH_3$ |
| | Formic acid (HCOOH) | HCOOH |
| | Formaldehyde (HCHO) | HCHO |
| | Carbon monoxide (CO) | CO |
| | Sulfur dioxide ($SO_2$) | $SO_2$ |
| | Primary organic aerosol | POA |
| | Primary nitrate | $PNO_3$ |
| | Primary sulfate | $PSO_4$ |
| | Other primary $PM_{2.5}$ | PMFINE |
| Ancillary data | Population density | PopDens |
| | Converted day of year | DOY |

to avoid over- or underestimation due to biased sample distribution. Then, the RF machine learning method was applied to the training datasets to develop the models for estimating ground-level PM concentrations.

### 3.1 Oversampling and subsampling

Many of the in situ observation data used in this study showed low concentrations, while there were a relatively small number of observations of high concentrations. This imbalance in samples could result in biased estimation with a significant underestimation of high-concentration data. Thus, over-/subsampling approaches were conducted for the training datasets to overcome the problem caused by the unbalanced samples (Table 3).

The oversampling approach is based on the assumption that the PM concentration of a training sample (i.e., at a pixel) is not significantly different from those of its neighboring pixels. The pixels within a circular window with a radius of three pixels (i.e., 37 pixels including the focus cell) were considered as potential neighboring pixels (see Supplement Fig. S1). Those 37 neighboring pixels were numbered based on the proximity to the center (i.e., the closer the pixel is to the center, the lower the number considering the direction from the focus). In order to perform oversampling, the intervals of 30 and 20 $\mu g\,m^{-3}$ were first applied to the $PM_{10}$ and $PM_{2.5}$ samples, respectively (i.e., 0–30, 30–60, …, 360–390, and > 390 $\mu g\,m^{-3}$ for $PM_{10}$ and 0–20, 20–40, …, 100–120, > 120 for $PM_{2.5}$). The second groups (i.e., 30–60 $\mu g\,m^{-3}$ for $PM_{10}$ and 20–40 $\mu g\,m^{-3}$ for $PM_{2.5}$) had the largest sample sizes, and thus the subsampling approach based on simple random sampling (i.e., 50 %) was applied to the second groups. For the other groups, we multiplied an integer value ranging from 1 to 37 by the sample size of each group to produce a more balanced sample distribution (i.e., the smaller the sample size, the larger the integer). Oversam-

**Table 3.** The number of samples for training, test, and hindcast validation datasets. The adjusted sample size for training data was determined through the over-/subsampling approaches.

| | Training dataset | | Test dataset | Hindcast validation dataset |
|---|---|---|---|---|
| | Original | Adjusted | | |
| $PM_{10}$ | 7919 | 14 201 | 1545 | 3906 |
| $PM_{2.5}$ | 3038 | 5738 | 776 | 1364 |

pling was then performed based on the order of the neighboring pixels. Input variables in the adjacent pixels of high-concentration samples were extracted with the corresponding target variables (i.e., $PM_{2.5}$ and $PM_{10}$) that were randomly perturbed within 5 % of the focus pixel concentrations. This oversampling approach can effectively reduce the underestimation of high PM concentrations that results from the small training sample size of high-concentration data.

### 3.2 Machine learning approach (random forest; RF)

RF is an ensemble model based on classification and regression trees (CART) with randomized node optimization and bootstrap aggregating (a.k.a. bagging; Breiman, 2001). RF generates numerous independent trees to overcome the limitations of a single-decision (or regression) tree method, such as the dependency on a single tree and the problem of overfitting the training data, resulting in better performance than single CARTs (Kim et al., 2015; Lee et al., 2016; Liu et al., 2018). A multitude of independent trees are ensembled to reach a solution by majority voting for classification or averaging for regression (e.g., Amani et al., 2017; Im et al., 2016; Latifi et al., 2018). RF provides information on how a variable contributes to model development using out-of-bag (OOB) data that are not used in training a model (Sonobe et al., 2017; Park et al., 2017). When a variable from OOB data is randomly permuted, the change in mean square error in percentage is calculated (Breiman, 2001). The larger the increase in the error for a variable, the more contributing the variable is. RF was applied to the training data to develop the models for estimating ground-level PM concentrations. The models were evaluated using the test and hindcast validation data.

### 3.3 Model evaluation

Accuracy assessment of the developed models were conducted using the test and hindcast validation datasets based on the following five metrics: coefficient of determination ($R^2$), RMSE, relative RMSE (rRMSE), mean bias (MB), and mean error (ME). rRMSE, MB, and ME are calculated as

$$\mathrm{rRMSE} = \frac{\mathrm{RMSE}}{\overline{y}} \times 100\,\%, \quad (1)$$

$$\mathrm{MB} = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i), \quad (2)$$

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}|f_i - y_i|, \quad (3)$$

where $y_i$ is the observed data, $\overline{y}$ is the mean of the observed data, $f_i$ is an estimated value, and $N$ is the number of observations. The rRMSE is the RMSE normalized by the mean value of observed data, which is useful for comparing results with different scales. The MB and ME are the averages of variation between the model-derived and observed values, with the exception that ME uses only absolute difference. The MB presents a tendency of overestimation or underestimation by a given model. The ME is the difference between observation and estimation (Boylan and Russell, 2006).

### 3.4 Comparison with other approaches

MODIS AOD is one of the widely used satellite-based aerosol products and has often been used to estimate PM concentrations. The developed RF models were compared with those using MODIS AOD instead of GOCI aerosol products. Unlike GOCI, MODIS only provides AOD with 3 km resolution (i.e., MYD04_3K) over land. AOD was used for developing MODIS-based models without incorporating other aerosol-related variables (i.e., AE, FMF, and SSA). In order to compare the performance between MODIS- and GOCI-based RF models, 50 % of the samples that were commonly included in both MODIS and GOCI datasets were used to develop the models, while the remaining samples were used to validate the models.

In addition, the ground-level PM concentrations predicted using the GOCI-based RF models were compared to the simulated and predicted results by GEOS-Chem and CMAQ models. The GEOS-Chem v10-01 was utilized with the Global Forecast System (GFS; produced by the National Centers for Environmental Prediction (NCEP)) as meteorological fields and the MIX Asian emission inventory was used as emissions. The nested domain for the GEOS-Chem simulation is 70–150° E and 15–55° N, which covers East Asia. The horizontal resolution of the nested model is $0.25° \times 0.3125°$. The boundary conditions for the nested model are from the GEOS-Chem global simulation at $2° \times 2.5°$ horizontal resolution. The CMAQ model version 4.7.1 was used to simulate the ground-level $PM_{10}$ and $PM_{2.5}$ concentrations. Meteorological fields simulated by the Weather Research and Forecasting (WRF) model and emission data from the SMOKE model were utilized to run the CMAQ model. The comparison among the GOCI-based model, GEOS-Chem, and CMAQ to in situ measurements was conducted using the hindcast validation dataset. For comparison to in situ measurements, the results from

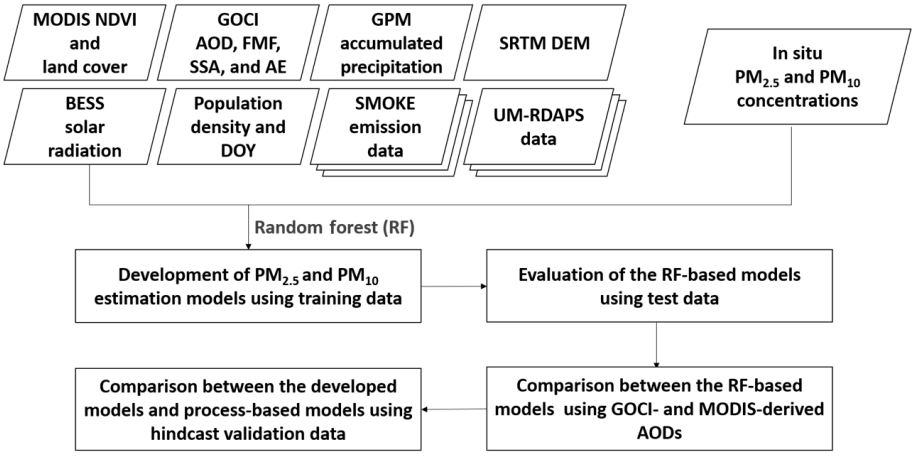


**Figure 2.** Process flow diagram of the estimation of ground-level particulate matter concentrations proposed in this study.

**Table 4.** Accuracy assessment results of the RF-based models for estimating PM concentrations using the test datasets during 2015–2016.

| | $R^2$ | RMSE[a] ($\mu g\,m^{-3}$) | rRMSE[b] (%) | MB[c] ($\mu g\,m^{-3}$) | ME[d] ($\mu g\,m^{-3}$) | Slope | Intercept |
|---|---|---|---|---|---|---|---|
| Model (with original training samples) | | | | | | | |
| $PM_{10}$ | 0.58 | 24.34 | 36.96 | −5.24 | 15.41 | 0.48 | 28.94 |
| $PM_{2.5}$ | 0.59 | 10.53 | 36.46 | −2.30 | 7.37 | 0.46 | 13.30 |
| Improved model (with balanced training samples) | | | | | | | |
| $PM_{10}$ | 0.78 | 17.08 | 25.94 | 2.93 | 12.78 | 0.78 | 17.16 |
| $PM_{2.5}$ | 0.73 | 8.25 | 28.58 | 1.71 | 6.18 | 0.77 | 8.30 |

[a] Root mean square error; [b] relative root mean square error; [c] mean bias; [d] mean error.

the GOCI-based models were resampled to the GEOS-Chem grid with $0.25° \times 0.3125°$ from January to September 2016 and to the CMAQ grids with $9\,km \times 9\,km$ for 2015–2016. The approach by van Donkelaar et al. (2010) that uses the ratio between the ground-level data and total column of AOD to satellite-based AOD (i.e., here GOCI AOD) using the vertical profile of AOD from GEOS-Chem was adopted to predict ground-level PM concentrations (i.e., GOCI-GEOS-Chem fused PM estimation).

## 4 Results and discussion

### 4.1 Performance of the RF models

The evaluation results of the developed models for estimating $PM_{10}$ and $PM_{2.5}$ concentrations using the test datasets over South Korea are presented in Table 4. The models (the improved models hereafter) based on the balanced training samples through over-/subsampling, resulted in $R^2$ values of 0.78 and 0.73 and RMSEs of 17.08 and $8.25\,\mu g\,m^{-3}$ for $PM_{10}$ and $PM_{2.5}$, respectively. There was a significant improvement in using the balanced training samples instead of the original samples (decrease in RMSE $\sim 30\,\%$ and rRMSE $\sim 10\,\%$). MB and ME also confirmed that the balanced samples improved the models estimating ground-level PM concentrations (Table 3; Fig. 3). In particular, high-concentration data (over $150\,\mu g\,m^{-3}$ for $PM_{10}$ and $50\,\mu g\,m^{-3}$ for $PM_{2.5}$) were well estimated by the improved models. The slopes of the trends were also improved from 0.46–0.48 to 0.77–0.78. The slopes were still lower than 1, which is due to the slight overestimation of low PM concentration data (Fig. 3). This significant improvement in the estimation performance was mainly due to the proposed sampling strategies in order to use more balanced training data. The use of the balanced training data resulted in the huge increase in the estimation accuracy of ground-level PM concentrations especially for high-concentration samples at the expense of a slight accuracy decrease for low concentrations.

Although it is not possible to directly compare the present results with those from other studies, the results from this study agreed well with those from recent literature that used machine learning approaches for estimating PM concentrations (Gupta and Christopher, 2009b; Wu et al., 2012; Li et al., 2017a; Yeganeh et al., 2017; Hu et al., 2017b; Chen et

**Table 5.** Seasonal variation in model performance for estimating particulate matter (PM) concentrations. Spring, summer, fall, and winter correspond to March to May, June to August, September to November, and December to February, respectively.

| | | $R^2$ | RMSE[a] ($\mu g\,m^{-3}$) | rRMSE[b] (%) | MB[c] ($\mu g\,m^{-3}$) | ME[d] ($\mu g\,m^{-3}$) | Slope | Intercept | Sample sizes ($N$) |
|---|---|---|---|---|---|---|---|---|---|
| $PM_{10}$ | Annual | 0.76 | 13.04 | 19.32 | 3.09 | 9.83 | 0.75 | 19.78 | 18 466 |
| | Spring | 0.74 | 13.07 | 17.77 | 3.08 | 9.98 | 0.70 | 25.06 | 13 132 |
| | Summer | 0.50 | 12.62 | 28.88 | 0.33 | 9.23 | 0.48 | 22.95 | 928 |
| | Fall | 0.77 | 16.61 | 26.69 | 7.76 | 11.81 | 0.87 | 15.76 | 1564 |
| | Winter | 0.87 | 12.78 | 19.22 | 3.71 | 9.20 | 0.87 | 12.29 | 2842 |
| $PM_{2.5}$ | Annual | 0.82 | 5.92 | 18.90 | 1.36 | 4.42 | 0.81 | 7.21 | 7188 |
| | Spring | 0.82 | 5.90 | 19.01 | 1.14 | 4.47 | 0.75 | 8.77 | 4510 |
| | Summer | 0.63 | 7.79 | 30.98 | 3.15 | 6.20 | 0.61 | 12.97 | 712 |
| | Fall | 0.85 | 8.12 | 27.50 | 3.89 | 6.53 | 0.88 | 7.30 | 961 |
| | Winter | 0.79 | 7.94 | 20.99 | 0.72 | 5.56 | 0.82 | 7.65 | 1005 |

[a] Root mean square error; [b] relative root mean square error; [c] mean bias; [d] mean error.

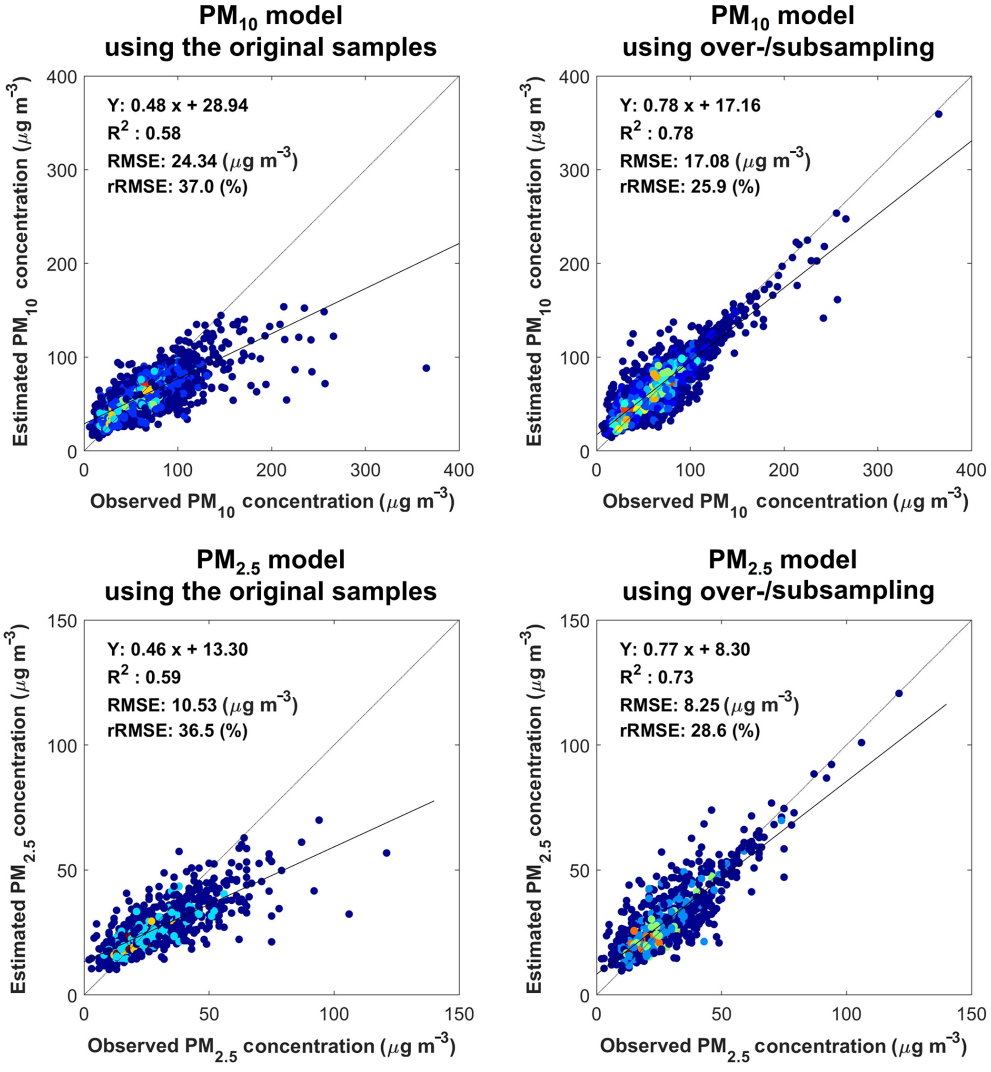


**Figure 3.** The model test results of daily $PM_{10}$ and $PM_{2.5}$ estimations. The color scheme from blue to red indicates the point density: the blue point means low density, while the red point shows high density.

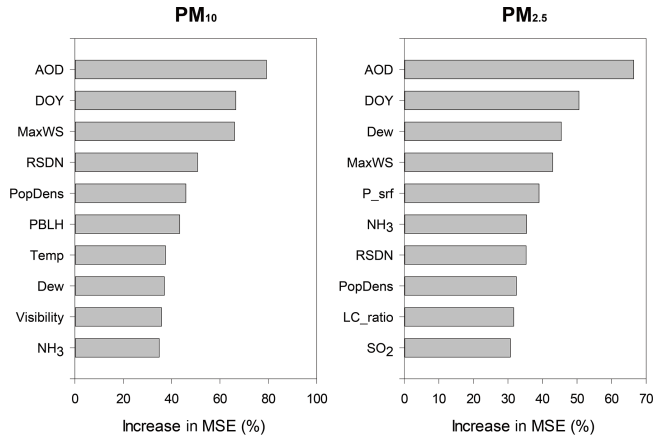


**Figure 4.** Variable importance of the top 10 input variables identified by the random forest models for estimating ground-level $PM_{10}$ and $PM_{2.5}$ concentrations.

al., 2018). Hu et al. (2017b) estimated surface $PM_{2.5}$ concentrations using RF, resulting in the cross-validation $R^2$ of 0.8 and RMSE of 2.83 $\mu g\,m^{-3}$. Similarly, Chen et al. (2018) compared three different methods (i.e., RF, the generalized additive model (GAM), and the nonlinear exposure–lag–response model (NEM)) to estimate surface $PM_{2.5}$ concentrations over China during 2014–2016. Their daily estimation results show cross-validation $R^2$ of 0.83, 0.55, and 0.51 for RF, GAM, and NEM, respectively, implying the robustness of machine learning compared to traditional statistical models. A geographically adjusted deep belief network (Geoi-DBN) was used to estimate $PM_{2.5}$ over China and showed a good correlation with observation data ($R^2 = 0.88$ and RMSE = 13.68 $\mu g\,m^{-3}$; Li et al., 2017a). The literature shows that empirical models using statistical and machine learning approaches often underestimate high PM concentrations (Wu et al., 2012; Li et al., 2017a). However, the RF-based models developed in our study has proved to be effective for modeling high ground-level PM concentrations.

In addition, the seasonal variation in model performance for 2015 and 2016 is shown in Table 5. The $R^2$ values for $PM_{10}$ estimations are the highest (0.87) in winter with an RMSE of 12.78 $\mu g\,m^{-3}$ and the lowest (0.50) in summer with an RMSE of 12.62 $\mu g\,m^{-3}$, as compared to $R^2$ values of 0.77 and 0.74 with RMSEs of 16.61 and 13.07 $\mu g\,m^{-3}$ in fall and spring, respectively. The summer season resulted in relatively high rRMSE for estimating ground-level PM concentrations compared to the other seasons. This is mainly because ground-level PM concentrations are typically low in summer in South Korea. The cloud contamination and the relatively small sample size in summer might lead to estimation errors (Shi et al., 2014; Sogacheva et al., 2017).

Figure 4 depicts the top 10 input variables that were identified as the most contributing variables by the improved RF models for estimating $PM_{10}$ and $PM_{2.5}$ concentrations. The results indicate that AOD, DOY, MaxWS (i.e., maximum wind speed), RSDN (i.e., solar radiation), and Dew (i.e., dew-point temperature) were commonly identified as contributing variables by the RF models to estimate both ground-level $PM_{10}$ and $PM_{2.5}$ concentrations. The AOD was identified as the most significant factor, which agreed well with the existing literature (Yu et al., 2017; Zang et al., 2017; Chen et al., 2018). Although most high PM concentration samples had high AOD values, some high-PM samples had low AOD values. Careful examination of the samples shows that there were Asian dust events at low altitudes in those cases, which were not effectively included in the AOD derived from satellite sensor systems. In other words, the satellite-derived AOD has a weak sensitivity in capturing aerosols at low altitudes (Choi et al., 2018). This could be an error source, implying that altitude information of such dust events can be used to further improve the models for estimating ground-level PM concentrations.

Some meteorological variables indicating the atmospheric conditions also contributed to the estimation of ground-level PM concentrations in the improved models. There is a relationship between solar radiation and aerosols in which solar radiation reaching the surface increases with decreasing aerosol concentration (Préndez et al., 1995; Hu et al., 2017a; Borlina and Rennó, 2017). Prior studies noted that there is an inverse relationship between wind speed and both $PM_{10}$ and $PM_{2.5}$ (Gupta et al., 2006; Maraziotis et al., 2008; Krynicka and Drzeniecka-Osiadacz, 2013). This relationship causes an increase in PM concentrations under low wind speed conditions but a decrease under high wind speed conditions, which is also confirmed in the present study. This means that atmospheric conditions such as air stagnation have significant impacts on surface PM concentrations. The results correspond to previous studies (e.g., You et al., 2015; Yeganeh et al., 2017; Hu et al., 2017b; Yu et al., 2017) showing that meteorological factors are strongly effective in improving PM estimation models. Interestingly, the anthropogenic factors such as LC_ratio (urban ratio), PopDens (population density), $NH_3$, and $SO_2$ were more important for $PM_{2.5}$ estimation than $PM_{10}$. This implies that the sources of $PM_{2.5}$ are mainly anthropogenic in South Korea (Moon et al., 2011; Ryou et al., 2018).

### 4.2 Spatial distribution of PM concentrations using the improved RF models

Figure 5 illustrates the spatial distribution of 2-year (2015–2016) averaged surface $PM_{10}$ and $PM_{2.5}$ concentrations at 1 km resolution with station-based in situ $PM_{10}$ and $PM_{2.5}$ concentrations over South Korea. The pixels that have concentration values for more than 5 % of the period (> 36 days for the 2 years) were used to produce the spatial distribution maps to secure the reliability of the distribution. The predicted $PM_{10}$ and $PM_{2.5}$ have similar spatial patterns with relatively high concentrations for urban areas especially around

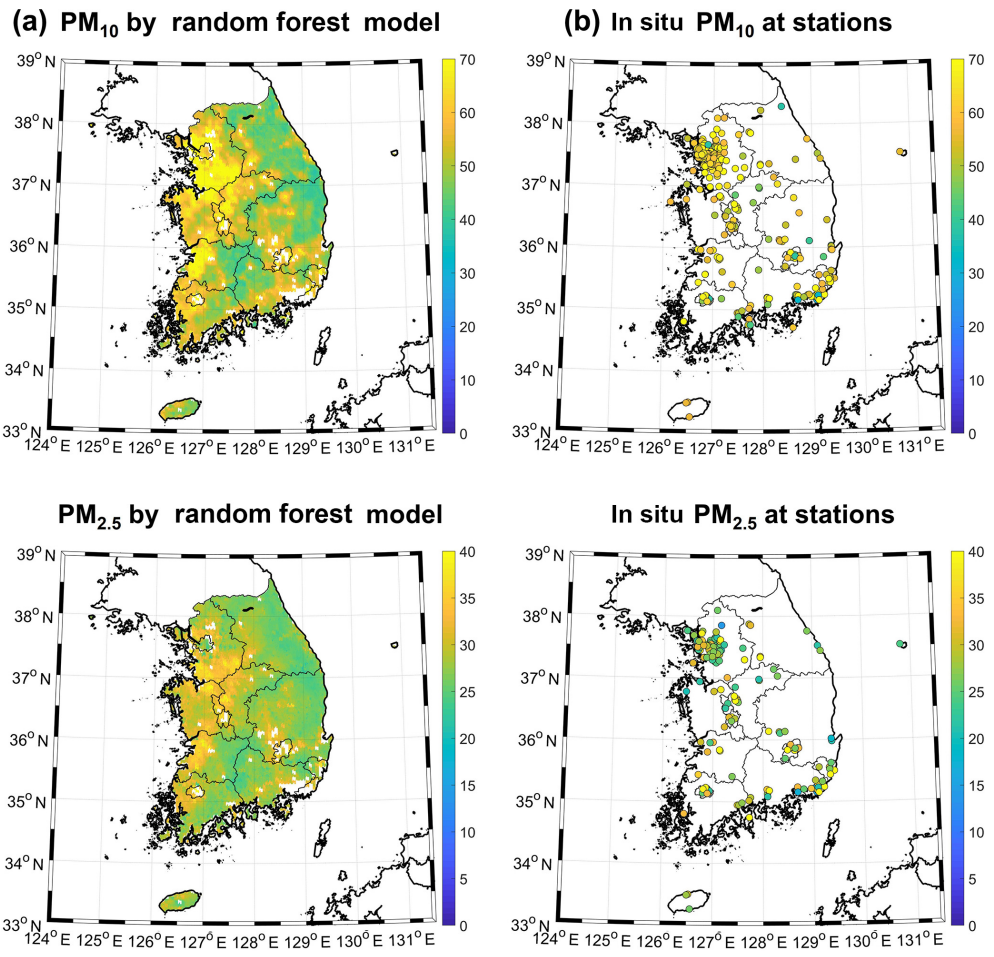


**Figure 5.** Maps of 2-year averaged particulate matter concentrations: $PM_{10}$ and $PM_{2.5}$ by the RF model **(a)**, and in situ $PM_{10}$ and $PM_{2.5}$ **(b)**.

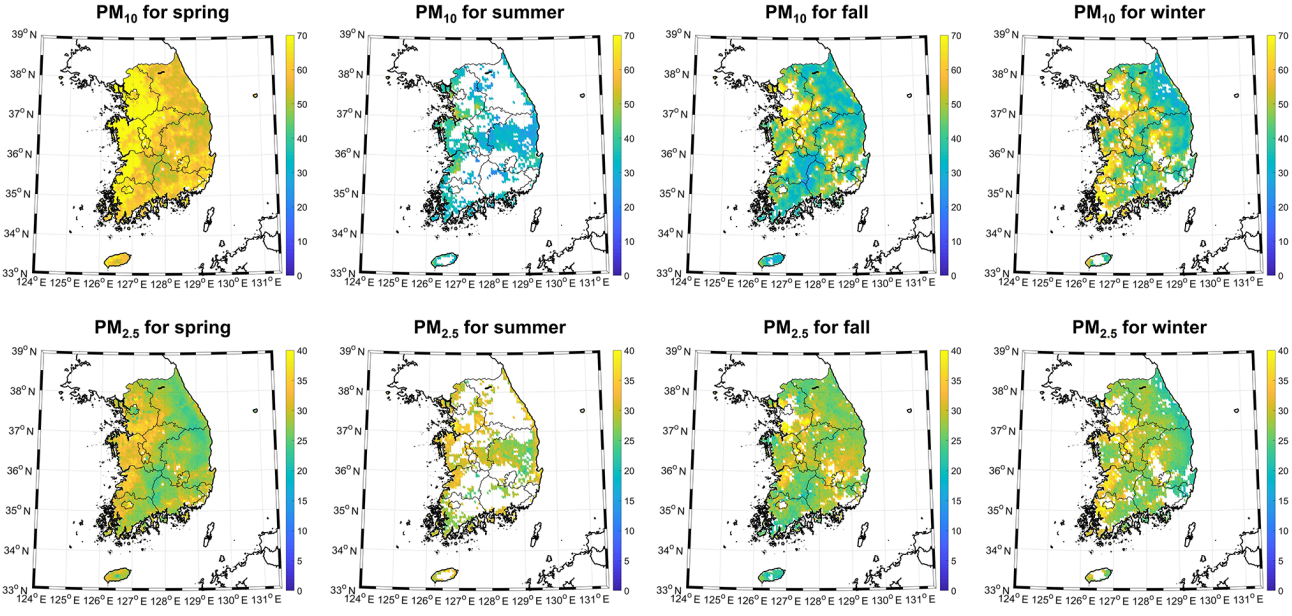


**Figure 6.** Spatial distributions of seasonal mean particulate matter concentrations (first row for $PM_{10}$ and second row for $PM_{2.5}$).

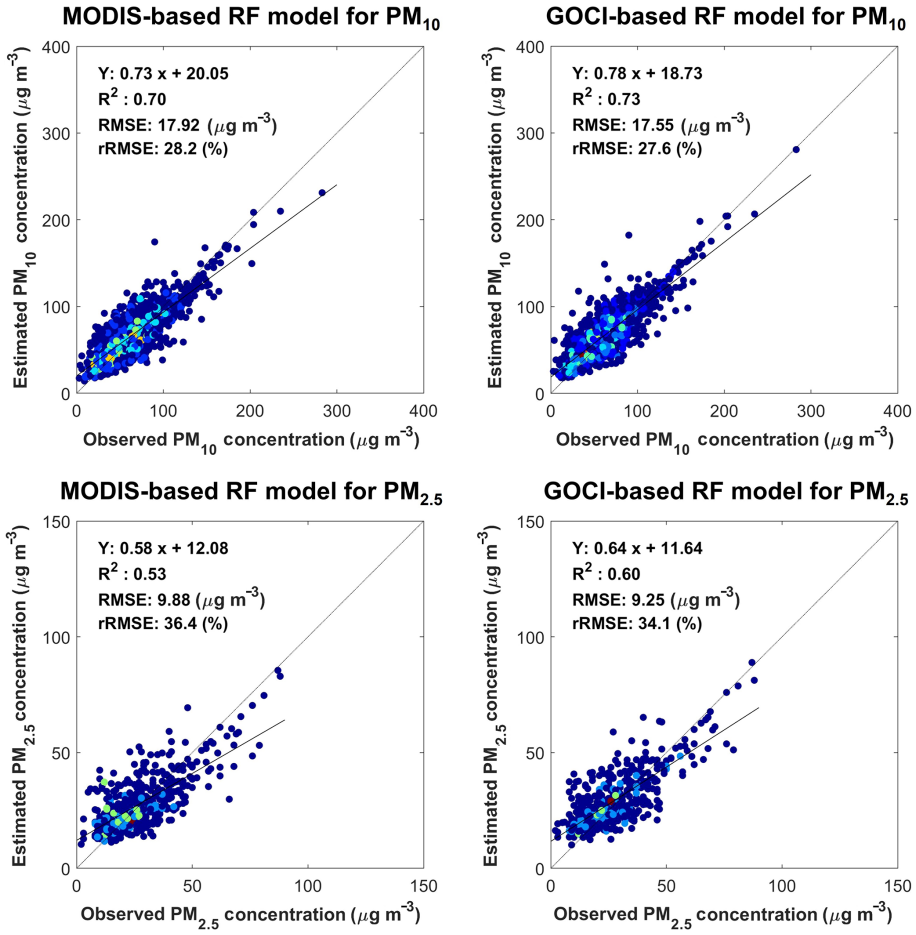


**Figure 7.** Scatterplots between the estimated and observed particulate matter concentrations when using MODIS- vs. GOCI-based models. The color scheme from blue to red indicates the point density: the blue point means low density, while the red point shows high density.

metropolitan areas and agree well with observed concentrations (Fig. 5).

The seasonal maps of $PM_{10}$ and $PM_{2.5}$ concentrations are also shown in Fig. 6. South Korea usually has the rainy season in June and July. For this reason, cloud contaminants are much more significant in the summer than in the other seasons, which resulted in many no-data pixels for the summer maps (Fig. 6). The ground-level PM concentrations in the spring and winter are much higher than in summer and fall for $PM_{10}$. The results agree well with the general seasonal patterns of $PM_{10}$ concentrations of South Korea, where PM concentrations are much higher in spring due to Asian dust inflow carried by westerly winds (Park and Shin, 2017). In addition, anthropogenic emissions generally increase PM concentrations in winter (Lu et al., 2011b; Li et al., 2016). The seasonal distribution of $PM_{2.5}$ concentrations is similar to that of $PM_{10}$. However, high concentrations were predominantly found in fall for $PM_{2.5}$. The cold Siberian high pressure might explain this. When warm air from the south flows into the study area, and while the force of the Siberian anticyclone stops, an inversion layer is formed. Then, PM is trapped because the atmospheric circulation becomes stagnant. Another reason may be the relative overestimation of $PM_{2.5}$ by the RF model in the fall season (Table 5). MB was greatest for the fall season among the four seasons, indicating an overestimation of $PM_{2.5}$. A more careful data configuration between training and test samples with larger sample size may mitigate such an overestimation.

### 4.3 Comparison of ground PM concentrations based on GOCI and MODIS AODs

The existing studies have generally used MODIS-derived AOD to estimate surface PM concentrations for various countries because of its global coverage and high quality (Remer et al., 2006; Gupta and Christopher, 2009a, b; Van Donkelaar et al., 2010; Wang et al., 2010; Chudnovsky et al., 2014; You et al., 2015; Hu et al., 2017b; Yu et al., 2017; He and Huang, 2018). In this section, the estimated ground-level $PM_{10}$ and $PM_{2.5}$ concentrations are compared based on GOCI AOD and MODIS AOD. Figure 7 displays the scatterplots showing the cross-validation results of the RF-based

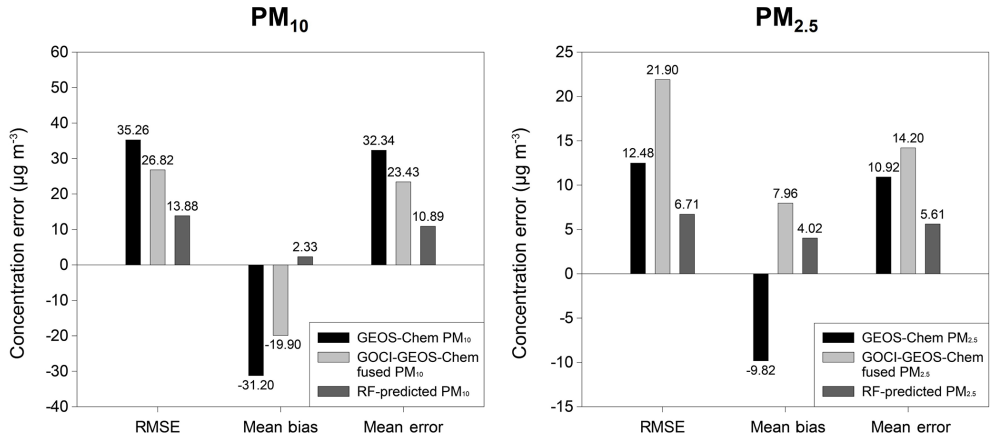


**Figure 8.** Comparison of the three models (i.e., GEOS-Chem based, GOCI-GEOS-Chem fused, and the present RF-based models) using the hindcast validation data for estimating particulate matter concentrations: $PM_{10}$ and $PM_{2.5}$ with root mean square error (RMSE), mean bias (MB), and mean error (ME).

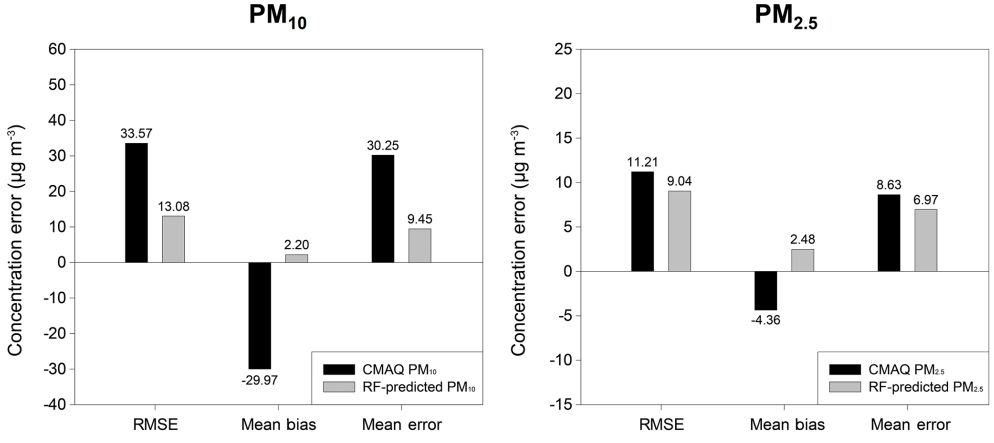


**Figure 9.** Comparison between the RF-based and CMAQ models using the hindcast validation data for estimating particulate matter concentrations: $PM_{10}$ and $PM_{2.5}$ with root mean square error (RMSE), mean bias (MB), and mean error (ME).

models using GOCI-derived and MODIS-derived AODs. Although there was no statistically significant difference between the two types of models through ANOVA tests, the GOCI-based RF models produced slightly better accuracy metrics (i.e., $R^2$, RMSE, and rRMSE) than MODIS-based RF models for estimating ground-level PM concentrations. When comparing ground PM concentrations to AODs derived from the two sensor data (i.e., MODIS and GOCI), GOCI-derived AOD showed slightly higher correlation with the ground PM concentrations than MODIS-derived ones (Supplement Fig. S2). Considering the advantages of GOCI as a geostationary satellite sensor (i.e., moderate spatial and temporal resolutions; eight times a day with a 6 km grid size of the aerosol product), it is very promising to use GOCI-derived products as input to PM estimation models. It should also be noted that GOCI-2, which has enhanced sensor specifications (i.e., 10 data collections per day at 3 km spatial resolution of the aerosol product) is planned to be launched in 2019.

### 4.4 Comparison with the process-based models

The RF-based models for estimating ground-level $PM_{10}$ and $PM_{2.5}$ concentrations were further compared with process-based models, i.e., GEOS-Chem and CMAQ. Figure 8 shows the comparison of the accuracy metrics of the three models: the GEOS-Chem simulated, GOCI-GEOS-Chem fused, and the RF-predicted PM concentrations using the hindcast validation datasets (Table 3). The GOCI-GEOS-Chem fused $PM_{10}$ concentrations have less errors than the GEOS-Chem simulated $PM_{10}$ concentration, which agrees well with the existing literature. However, both tend to significantly underestimate the ground-level $PM_{10}$ concentrations when compared to the proposed RF models. Consequently, the proposed RF models have the lowest RMSE, MB, and ME among those models. Although the results of

GOCI-GEOS-Chem fused $PM_{2.5}$ showed that $R^2$ (GEOS-Chem $PM_{2.5}$: 0.00; GOCI-GEOS-Chem fused $PM_{2.5}$: 0.14) and slope (GEOS-Chem $PM_{2.5}$: −0.02; GOCI-GEOS-Chem fused $PM_{2.5}$: 1.41) improved more than those of GEOS-Chem $PM_{2.5}$, the RMSE and ME of the fused model were higher than the GEOS-Chem model because the fused model overestimated PM concentrations. The RF models also produced a better performance than CMAQ for estimating both $PM_{10}$ and $PM_{2.5}$ concentrations (Fig. 9). Similar to the GEOS-Chem models, CMAQ tends to underestimate PM concentrations showing a large negative MB value.

## 5 Conclusions

In this study, machine learning (i.e., RF) based models were developed to estimate ground-level $PM_{10}$ and $PM_{2.5}$ concentrations through the synergistic use of satellite data and model output over South Korea. The RF-based models developed using the balanced training samples produced good performance resulting in $R^2$ values of 0.78 and 0.73 and RMSEs of 17.08 and 8.25 $\mu g\,m^{-3}$ for $PM_{10}$ and $PM_{2.5}$, respectively. In particular, the proposed models estimated high PM concentrations well. GOCI-derived AOD was identified as the most significant input variable for estimating ground-level PM concentrations. A few meteorological variables such as MaxWS, RSDN, and dew-point temperature were also revealed as contributing variables. In addition, the anthropogenic factors such as urban ratio, population density, and emission of $SO_2$ and $NH_3$ were considered significant for estimating $PM_{2.5}$ concentrations. Two-year and seasonally averaged maps of ground-level PM concentrations agree with spatiotemporal patterns of PM concentrations reported in the literature.

The proposed RF models were also compared to the two process-based models (GEOS-Chem and CMAQ) using the hindcast validation data. When GOCI-derived AOD was incorporated with the GEOS-Chem data, the estimation of PM concentrations improved. However, the incorporated approach still underestimated high concentrations when compared to the proposed RF models. Similar results were found for the comparison between the RF models and CMAQ, which implies the robustness of the proposed approach.

Although the proposed models performed better than the existing models, there are several ways to further improve the proposed models, which deserve further investigation. First, more input variables, especially those that are related to vertical information of AOD, can be used to improve the models. In addition, other sophisticated approaches such as deep learning could be utilized to improve the estimation accuracy for ground-level PM concentrations. Although only 2-year data were used in this study, longer archives can be used to further refine the models. The synergistic use of forthcoming geostationary satellite series of Geostationary – Korea Multi-Purpose Satellite-2A (GEO-KOMPSAT-2A; GK-2A) with Advanced Meteorological Imager (AMI) and GK-2B with GOCI-II and Geostationary Environment Monitoring Spectrometer (GEMS) sensors will provide more accurate aerosol information with higher spatial and temporal resolutions than those of GOCI. Such a synergy is likely to improve the estimation of ground-level PM concentrations in the near future.

*Data availability.* Data are available upon request to the corresponding author.

*Supplement.* The supplement related to this article is available online at: https://doi.org/10.5194/acp-19-1097-2019-supplement.

*Author contributions.* SP and MS contributed equally to the paper. SP and MS led the manuscript writing and contributed to the research design and data analysis. JI supervised this study, contributed to the research design and manuscript writing, and served as the corresponding author. CS contributed to the discussion of the results and manuscript writing. MC, JhK, SL, and RP contributed to data processing and the discussion of the results. JiK, DL, and SK contributed to the data sharing and discussion of the results.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* This study was supported by a grant from the National Institute of Environmental Research (NIER), funded by the Ministry of Environment (MOE) of the Republic of Korea (NIER-2017-01-02-063), the Space Technology Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning (NRF-2017M1A3A3A02015981), and the National Strategic Project-Fine Particle of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT), the Ministry of Environment (ME), and the Ministry of Health and Welfare (MOHW) (NRF-2017M3D8A1092021). MC's work was undertaken as a private enterprise and not in the author's capacity as an employee of the Jet Propulsion Laboratory, California Institute of Technology.

Edited by: Michael Schulz
Reviewed by: three anonymous referees

## References

AirKorea: https://www.airkorea.or.kr/, last access: 24 January 2019.

Amani, M., Salehi, B., Mahdavi, S., Granger, J., and Brisco, B.: Wetland classification in Newfoundland and Labrador using multi-source SAR and optical data integration, GISci. Remote Sens., 54, 779–796, 2017.

Baek, B. H., Seppanen, C., and Houyoux, M.: SMOKE v2. 5 User's manual, https://www.cmascenter.org/smoke/documentation/2.5/html/, last access: 24 January 2019.

Bartell, S. M., Longhurst, J., Tjoa, T., Sioutas, C., and Delfino, R. J.: Particulate air pollution, ambulatory heart rate variability, and cardiac arrhythmia in retirement community residents with coronary artery disease, Environ. Health Persp., 121, 1135–1141, https://doi.org/10.1289/ehp.1205914, 2013.

Borlina, C. S. and Rennó, N. O.: The Impact of a Severe Drought on Dust Lifting in California's Owens Lake Area, Sci. Rep., 7, 1784, https://doi.org/10.1038/s41598-017-01829-7, 2017.

Boylan, J. W. and Russell, A. G.: PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models, Atmos. Environ., 40, 4946–4959, 2006.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Brokamp, C., Jandarov, R., Hossain, M., and Ryan, P.: Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model, Environ. Sci. Technol., 52, 4173–4179, 2018.

Chen, G., Li, S., Knibbs, L. D., Hamm, N., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J., and Guo, Y.: A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information, Sci. Total Environ., 636, 52–60, 2018.

Choi, M.: Retrieval of aerosol optical properties from GOCI: Algorithm Improvement, Analysis and Application to PM (Doctoral dissertation), Graduate School, Yonsei University, Seoul, 2017.

Choi, M., Kim, J., Lee, J., Kim, M., Park, Y.-J., Holben, B., Eck, T. F., Li, Z., and Song, C. H.: GOCI Yonsei aerosol retrieval version 2 products: an improved algorithm and error analysis with uncertainty estimation from 5-year validation over East Asia, Atmos. Meas. Tech., 11, 385–408, https://doi.org/10.5194/amt-11-385-2018, 2018.

Chudnovsky, A. A., Koutrakis, P., Kloog, I., Melly, S., Nordio, F., Lyapustin, A., Wang, Y., and Schwartz, J.: Fine particulate matter predictions using high resolution Aerosol Optical Depth (AOD) retrievals, Atmos. Environ., 89, 189–198, 2014.

Davies, T., Cullen, M. J., Malcolm, A. J., Mawson, M., Staniforth, A., White, A., and Wood, N.: A new dynamical core for the Met Office's global and regional modelling of the atmosphere, Q. J. Roy. Meteor. Soc., 131, 1759–1782, 2005.

Earthdata: https://search.earthdata.nasa.gov/, last access: 24 January 2019.

Environmental Ecology Lab at Seoul National University: http://environment.snu.ac.kr/bess_rad/, last access: 24 January 2019.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., and Roth, L.: The shuttle radar topography mission, Rev. Geophys., 45, RG2004, https://doi.org/10.1029/2005RG0001832007, 2007.

Forkuor, G., Dimobe, K., Serme, I, and Tondoh, J.: Landsat-8 vs. Sentinel-2: examining the added value of sentinel-2's red-edge bands to land-use and land-cover mapping in Burkina Faso, GISci. Remote Sens., 55, 331–354, 2018.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X.: MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets, Remote Sens. Environ., 114, 168–182, 2010.

Gao, M., Carmichael, G. R., Saide, P. E., Lu, Z., Yu, M., Streets, D. G., and Wang, Z.: Response of winter fine particulate matter concentrations to emission and meteorology changes in North China, Atmos. Chem. Phys., 16, 11837–11851, https://doi.org/10.5194/acp-16-11837-2016, 2016.

GES DISC: https://mirador.gsfc.nasa.gov/, last access: 24 January 2019.

Gupta, A., Nag, S., and Mukhopadhyay, U.: Characterisation of $PM_{10}$, $PM_{2.5}$ and benzene soluble organic fraction of particulate matter in an urban area of Kolkata, India, Environ. Monit. Assess., 115, 205–222, 2006.

Gupta, P. and Christopher, S. A.: Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach, J. Geophys. Res.-Atmos., 114, D14205, https://doi.org/10.1029/2008JD011496, 2009a.

Gupta, P. and Christopher, S. A.: Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach, J. Geophys. Res.-Atmos., 114, D20205, https://doi.org/https://doi.org/10.1029/2008JD01149, 2009b.

He, Q. and Huang, B.: Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling, Remote Sens. Environ., 206, 72–83, 2018.

Hu, B., Zhao, X., Liu, H., Liu, Z., Song, T., Wang, Y., Tang, L., Xia, X., Tang, G., and Ji, D.: Quantification of the impact of aerosol on broadband solar radiation in North China, Sci. Rep., 7, 44851, https://doi.org/10.1038/srep44851, 2017a.

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., and Liu, Y.: Estimating $PM_{2.5}$ Concentrations in the Conterminous United States Using the Random Forest Approach, Environ. Sci. Technol., 51, 6936–6944, 2017b.

Huffman, G. J., Bolvin, D. T., and Nelkin, E. J.: Integrated MultisatellitE Retrievals for GPM (IMERG) technical documentation, NASA/GSFC Code, 612, 47, https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_doc.05.pdf (last access: 24 Janaury 2019), 2015.

Im, J., Jensen, J., Coleman, M., and Nelson, E.: Hyperspectral remote sensing analysis of short rotation woody crops grown with controlled nutrient and irrigation treatments, Geocarto Int., 24, 293–312, 2009.

Im, J., Park, S., Rhee, J., Baik, J., and Choi, M.: Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches, Environ. Earth Sci., 75, 1120, https://doi.org/10.1007/s12665-016-5917-6, 2016.

Jang, E., Im, J., Park, G.-H., and Park, Y.-G.: Estimation of fugacity of carbon dioxide in the East Sea using in situ measurements and Geostationary Ocean Color Imager satellite data, Remote Sens., 9, 821, https://doi.org/10.3390/rs9080821, 2017.

Jerrett, M., Turner, M. C., Beckerman, B. S., Pope III, C. A., van Donkelaar, A., Martin, R. V., Serre, M., Crouse, D., Gapstur, S. M., and Krewski, D.: Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates, Environ. Health Persp., 125, 55–559, 2017.

Ke, Y., Im, J., Park, S., and Gong, H.: Downscaling of MODIS One kilometer evapotranspiration using Landsat-8 data and machine learning approaches, Remote Sens., 8, 215, https://doi.org/10.3390/rs8030215, 2016.

Kim, M., Im, J., Han, H., Kim, J., Lee, S., Shin, M., and Kim, H.: Landfast sea ice monitoring using multi sensor fusion in the Antarctic, GISci. Remote Sens., 52, 239–256, 2015.

Koo, Y.-S., Kim, S.-T., Cho, J.-S., and Jang, Y.-K.: Performance evaluation of the updated air quality forecasting system for Seoul predicting $PM_{10}$, Atmos.Environ., 58, 56–69, 2012.

Krynicka, J. and Drzeniecka-Osiadacz, A.: Analysis of Variability in $PM_{10}$ Concentration in the Wrocław Agglomeration, Pol. J. Environ. Stud., 22, 1091–1099, 2013.

Latifi, H., Dahms, T., Beudert, B., Heurich, M., Kubert, C., and Dech, S.: Synthetic RapidEye data used for the detection of area-based spruce tree mortality induced by bark beetles, GISci. Remote Sens., 55, 839–959, 2018.

Lee, S., Im, J., Kim, J., Kim, M., Shin, M., Kim, H.-c., and Quackenbush, L. J.: Arctic sea ice thickness estimation from CryoSat-2 satellite data using machine learning-based lead detection, Remote Sens., 8, 698, 2016.

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C.: The Collection 6 MODIS aerosol products over land and ocean, Atmos. Meas. Tech., 6, 2989–3034, https://doi.org/10.5194/amt-6-2989-2013, 2013.

Li, K., Liao, H., Mao, Y., and Ridley, D. A.: Source sector and region contributions to concentration and direct radiative forcing of black carbon in China, Atmos. Environ., 124, 351–366, 2016.

Li, R., Gong, J., Chen, L., and Wang, Z.: Estimating ground-level $PM_{2.5}$ using fine-resolution satellite data in the megacity of Beijing, China, Aerosol Air Qual. Res., 15, 1347–1356, 2015.

Li, T., Shen, H., Yuan, Q., Zhang, X., and Zhang, L.: Estimating Ground-Level $PM_{2.5}$ by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach, Geophys. Res. Lett., 44, 11985–11993, 2017a.

Li, T., Shen, H., Zeng, C., Yuan, Q., and Zhang, L.: Point-surface fusion of station measurements and satellite observations for mapping $PM_{2.5}$ distribution in China: Methods and assessment, Atmos. Environ., 152, 477–489, 2017b.

Liu, T., Im, J., and Quackenbush, L. J.: A novel transferable individual tree crown delineation model based on Fishing Net Dragging and boundary classification, ISPRS J. Photogramm., 110, 34–47, 2015.

Liu, T., Abd-Elrahman, A., Morton, J., and Wilhelm, V.: Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system, GISci. Remote Sens., 55, 243–264, 2018.

Liu, Y., Sarnat, J. A., Kilaru, V., Jacob, D. J., and Koutrakis, P.: Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing, Environ. Sci. Technol., 39, 3269–3278, 2005.

Liu, Y., Franklin, M., Kahn, R., and Koutrakis, P.: Using aerosol optical thickness to predict ground-level $PM_{2.5}$ concentrations in the St. Louis area: A comparison between MISR and MODIS, Remote Sens. Environ., 107, 33–44, 2007.

Liu, Y., Paciorek, C. J., and Koutrakis, P.: Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information, Environ. Health Persp., 117, 886–892, 2009.

Lu, Z., Im, J., and Quackenbush, L.: A volumetric approach to population estimation using Lidar remote sensing, Photogramm. Eng. Rem. S., 77, 1145–1156, 2011a.

Lu, Z., Zhang, Q., and Streets, D. G.: Sulfur dioxide and primary carbonaceous aerosol emissions in China and India, 1996–2010, Atmos. Chem. Phys., 11, 9839–9864, https://doi.org/10.5194/acp-11-9839-2011, 2011b.

Lv, B., Hu, Y., Chang, H. H., Russell, A. G., Cai, J., Xu, B., and Bai, Y.: Daily estimation of ground-level $PM_{2.5}$ concentrations at 4 km resolution over Beijing-Tianjin-Hebei by fusing MODIS AOD and ground observations, Sci. Total Environ., 580, 235–244, 2017.

Maraziotis, E., Sarotis, L., Marazioti, C., and Marazioti, P.: Statistical analysis of inhalable ($PM_{10}$) and fine particles ($PM_{2.5}$) concentrations in urban region of Patras, Greece, Global Nest J., 10, 123–131, 2008.

Moon, K.-J., Park, S.-M., Park, J.-S., Song, I.-H., Jang, S.-K., Kim, J.-C., and Lee, S.-J.: Chemical Characteristics and Source Apportionment of $PM_{2.5}$ in Seoul Metropolitan Area in 2010, J. Korean Soc. Atmos. Environ., 27, 711–722, 2011.

Park, M. E., Song, C. H., Park, R. S., Lee, J., Kim, J., Lee, S., Woo, J.-H., Carmichael, G. R., Eck, T. F., Holben, B. N., Lee, S.-S., Song, C. K., and Hong, Y. D.: New approach to monitor transboundary particulate pollution over Northeast Asia, Atmos. Chem. Phys., 14, 659–674, https://doi.org/10.5194/acp-14-659-2014, 2014.

Park, S. and Shin, H.: Analysis of the Factors Influencing $PM_{2.5}$ in Korea: Focusing on Seasonal Factors, J. Environ. Pol. Admin., 25, 227–248, 2017.

Park, S., Im, J., Park, S., and Rhee, J.: Drought monitoring using high resolution soil moisture through machine learning approaches over the Korean peninsula, Agr. Forest Meteorol., 237, 257–269, 2017.

Pham, T., Yoshino, K., and Bui, D.: Biomass estimation of Sonneratia caseolaris (l.) Engler at a coastal area of Hai Phong city (Vietnam) using ALOS-2 PALSAR imagery and GIS-based multilayer perceptron neural networks, GISci. Remote Sens., 54, 329–353, 2017.

Pope III, C. A., Ezzati, M., and Dockery, D. W.: Fine-particulate air pollution and life expectancy in the United States, New Engl. J. Med., 360, 376–386, 2009.

Préndez, M. M., Egido, M., Tomas, C., Seco, J., Calvo, A., and Romero, H.: Correlation between solar radiation and total syspended particulate matter in Santiago, Chile – Preliminary results, Atmos. Environ., 29, 1543–1551, 1995.

Remer, L. A., Tanre, D., Kaufman, Y. J., Levy, R., and Mattoo, S.: Algorithm for remote sensing of tropospheric aerosol from MODIS: Collection 005, National Aeronautics and Space Administration, 1490, https://modis-atmosphere.gsfc.nasa.gov/sites/default/files/ModAtmo/ATBD_MOD04_C005_rev2_0.pd (last access: 24 January 2019), 2006.

Richardson, H., Hill, D., Denesiuk, D., and Fraser, L.: A comparison of geographic datasets and field measurements to model soil carbon using random forests and stepwise regressions (British Columbia, Canada), GISci. Remote Sens., 54, 573–591, 2017.

Ryou, H., Heo, J., and Kim, S.-Y.: Source apportionment of $PM_{10}$ and $PM_{2.5}$ air pollution, and possible impacts of study characteristics in South Korea, Environ. Pollut., 240, 963–972, 2018.

Ryu, Y., Jiang, C., Kobayashi, H., and Detto, M.: MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5 km resolution from 2000, Remote Sens. Environ., 204, 812–825, 2018.

SGIS: https://sgis.kostat.go.kr/, last access: 24 January 2019.

Shi, Y., Zhang, J., Reid, J. S., Liu, B., and Hyer, E. J.: Critical evaluation of cloud contamination in the MISR aerosol products using MODIS cloud mask products, Atmos. Meas. Tech., 7, 1791–1801, https://doi.org/10.5194/amt-7-1791-2014, 2014.

Sogacheva, L., Kolmonen, P., Virtanen, T. H., Rodriguez, E., Saponaro, G., and de Leeuw, G.: Post-processing to remove residual clouds from aerosol optical depth retrieved using the Advanced Along Track Scanning Radiometer, Atmos. Meas. Tech., 10, 491–505, https://doi.org/10.5194/amt-10-491-2017, 2017.

Solano, R., Didan, K., Jacobson, A., and Huete, A.: MODIS vegetation index user's guide (MOD13 series), Vegetation Index and Phenology Lab, The University of Arizona, 1–38, 2010.

Sonobe, R., Yamaya, Y., Tani, H., Wang, X., Kobayashi, N., and Mochizuki, K.: Assessing the suitability of data from Sentinel-1A and 2A for crop classification, GISci. Remote Sens., 54, 918–938, 2017.

Stolwijk, A., Straatman, H., and Zielhuis, G.: Studying seasonality by using sine and cosine functions in regression analysis, J. Epidemiol. Commun. H., 53, 235–238, 1999.

USGS EarthExplorer: https://earthexplorer.usgs.gov/, last access: 24 January 2019.

Van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., and Villeneuve, P. J.: Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application, Environ. Health Persp., 118, 847–855, 2010.

Van Donkelaar, A., Martin, R. V., Brauer, M., and Boys, B. L.: Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter, Environ. Health Persp., 123, 3762–3772, 2015.

van Zelm, R., Preiss, P., van Goethem, T., Van Dingenen, R., and Huijbregts, M.: Regionalized life cycle impact assessment of air pollution on the global scale: damage to human health and vegetation, Atmos. Environ., 134, 129–137, 2016.

Wang, Z., Chen, L., Tao, J., Zhang, Y., and Su, L.: Satellite-based estimation of regional particulate matter (PM) in Beijing using vertical-and-RH correcting method, Remote Sens. Environ., 114, 50–63, 2010.

Wu, J., Yao, F., Li, W., and Si, M.: VIIRS-based remote sensing estimation of ground-level $PM_{2.5}$ concentrations in Beijing–Tianjin–Hebei: A spatiotemporal statistical model, Remote Sens. Environ., 184, 316–328, 2016.

Wu, Y., Guo, J., Zhang, X., Tian, X., Zhang, J., Wang, Y., Duan, J., and Li, X.: Synergy of satellite and ground based observations in estimation of particulate matter in eastern China, Sci. Total Environ., 433, 20–30, 2012.

Xu, J.-W., Martin, R. V., van Donkelaar, A., Kim, J., Choi, M., Zhang, Q., Geng, G., Liu, Y., Ma, Z., Huang, L., Wang, Y., Chen, H., Che, H., Lin, P., and Lin, N.: Estimating ground-level $PM_{2.5}$ in eastern China using aerosol optical depth determined from the GOCI satellite instrument, Atmos. Chem. Phys., 15, 13133-13144, https://doi.org/10.5194/acp-15-13133-2015, 2015a.

Xu, L., Guo, H., Boyd, C. M., Klein, M., Bougiatioti, A., Cerully, K. M., Hite, J. R., Isaacman-VanWertz, G., Kreisberg, N. M., and Knote, C.: Effects of anthropogenic emissions on aerosol formation from isoprene and monoterpenes in the southeastern United States, P. Natl. Acad. Sci. USA, 112, 37–42, 2015b.

Yeganeh, B., Hewson, M. G., Clifford, S., Knibbs, L. D., and Morawska, L.: A satellite-based model for estimating $PM_{2.5}$ concentration in a sparsely populated environment using soft computing techniques, Environ. Modell. Softw., 88, 84–92, 2017.

Yoo, C., Im, J., Park, S., and Quackenbush, L. J.: Estimation of daily maximum and minimum air temperatures in urban landscapes using MODIS time series satellite data, ISPRS J. Photogramm., 137, 149–162, 2018.

Yoo, S., Im, J., and Wagner, J.: Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY, Landscape Urban Plan., 107, 293–306, 2012.

You, W., Zang, Z., Zhang, L., Li, Z., Chen, D., and Zhang, G.: Estimating ground-level $PM_{10}$ concentration in northwestern China using geographically weighted regression based on satellite AOD combined with CALIPSO and MODIS fire count, Remote Sens. Environ., 168, 276–285, 2015.

Yu, W., Liu, Y., Ma, Z., and Bi, J.: Improving satellite-based $PM_{2.5}$ estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting, Sci. Rep., 7, 7048, https://doi.org/10.1038/s41598-017-07478-0, 2017.

Zang, Z., Wang, W., You, W., Li, Y., Ye, F., and Wang, C.: Estimating ground-level $PM_{2.5}$ concentrations in Beijing, China using aerosol optical depth and parameters of the temperature inversion layer, Sci. Total Environ., 575, 1219–1227, 2017.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., and Zhang, M.: Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm, Atmos. Environ., 155, 129-139, 2017.

Zhao, S., Yu, Y., Yin, D., He, J., Liu, N., Qu, J., and Xiao, J.: Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from China National Environmental Monitoring Center, Environ. Int., 86, 92–106, 2016.