ISPRS Journal of Photogrammetry and Remote Sensing 183 (2022) 253–268

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Improved retrievals of aerosol optical depth and fine mode fraction from GOCI geostationary satellite data using machine learning over East Asia

Yoojin Kang [1], Miae Kim [1], Eunjin Kang, Dongjin Cho, Jungho Im [*]

Department of Urban & Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea

ARTICLE INFO

Keywords:
Aerosol optical depth
Fine mode fraction
Geostationary Ocean Color Imager
Machine learning
Shapley Additive exPlanations values

ABSTRACT

Aerosol Optical Depth (AOD) and Fine Mode Fraction (FMF) are important information for air quality research. Both are mainly obtained from satellite data based on a radiative transfer model, which requires heavy computation and has uncertainties. We proposed machine learning-based models to estimate AOD and FMF directly from Geostationary Ocean Color Imager (GOCI) reflectances over East Asia. Hourly AOD and FMF were estimated for 00–07 UTC at a spatial resolution of 6 km using the GOCI reflectances, their channel differences (with 30-day minimum reflectance), solar and satellite viewing geometry, meteorological data, geographical information, and the Day Of the Year (DOY) as input features. Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) machine learning approaches were applied and evaluated using random, spatial, and temporal 10-fold cross-validation with ground-based observation data. LightGBM (R^2 = 0.89–0.93 and RMSE = 0.071–0.091 for AOD and R^2 = 0.67–0.81 and RMSE = 0.079–0.105 for FMF) and RF (R^2 = 0.88–0.92 and RMSE = 0.080–0.095 for AOD and R^2 = 0.59–0.76 and RMSE = 0.092–0.118 for FMF) agreed well with the *in-situ* data. The machine learning models showed much smaller errors when compared to GOCI-based Yonsei aerosol retrieval and the Moderate Resolution Imaging Spectroradiometer Dark Target and Deep Blue algorithms. The Shapley Additive exPlanations values (SHAP)-based feature importance result revealed that the 412 nm band (i.e., ch01) contributed most in both AOD and FMF retrievals. Relative humidity and air temperature were also identified as important factors especially for FMF, which suggests that considering meteorological conditions helps improve AOD and FMF estimation. Besides, spatial distribution of AOD and FMF showed that using the channel difference features to indirectly consider surface reflectance was very helpful for AOD retrieval on bright surfaces.

1. Introduction

Atmospheric aerosols are liquid and solid particles suspended in the atmosphere. They come from a variety of natural and anthropogenic sources, which include the emissions of primary particulate matter or the formation of secondary particulate matter from gaseous precursors. Sea salt, most mineral dust, and primary biological aerosol particles (PBAPs) mainly originate from natural sources, while black carbon (BC), sulphate, nitrate and ammonium generally come from anthropogenic sources (Boucher et al., 2013). It is well known that aerosols can negatively affect human health and even terrestrial and marine ecosystems such as changes in vegetation coverage and plankton ecosystems (Pöschl, 2005; Rap et al., 2018; Unnithan and Gnanappazham, 2020). Atmospheric aerosols also cause changes in radiative forcing as they interact with radiation and clouds, affecting the Earth's radiation budget (Boucher et al., 2013). Therefore, it is necessary to understand the optical and physical properties of aerosols and their spatial distribution for accurate estimation of aerosols.

Aerosol optical depth (AOD) is a measure of the extinction of the solar radiation by aerosol particles in the atmosphere. It represents solar radiation reduction by aerosol particles, as they reflect, absorb or scatter sunlight (Boucher et al., 2013; Shin et al., 2020). AOD patterns are therefore controlled by aerosol patterns, and spatially and temporally vary with weather and geographical conditions (Della Ceca et al., 2018). AOD has been useful information in climate change and air pollution research (Martins et al., 2018). It has been used as a major predictor for estimating ground-level particulate matters (Kim et al., 2021; Yao et al., 2019; Park et al., 2020). In a global scale, effective radiative forcings are

* Corresponding author.
E-mail address: ersgis@unist.ac.kr (J. Im).
[1] These authors equally contributed to the paper.

https://doi.org/10.1016/j.isprsjprs.2021.11.016
Received 1 September 2021; Received in revised form 28 October 2021; Accepted 19 November 2021
Available online 30 November 2021

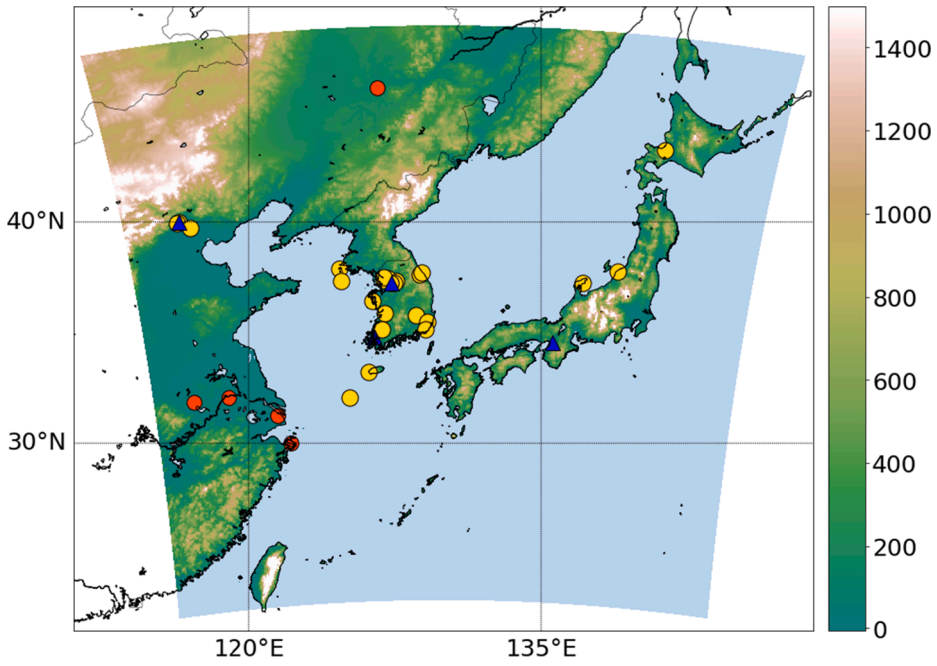

Fig. 1. Study area with *in-situ* observation sites. The AERONET stations are shown in yellow circles and the SONET stations in red circles. The independent test stations are marked as blue triangles. The background image is surface elevation (m) from the Japanese Aerospace Exploration Agency (JAXA) ALOS World 3D 30 m (AW3D30) Digital Surface Model (DSM). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

used to quantify the effect of aerosols on radiation and cloud, which are radiative forcing from aerosol-radiation interactions (ERFari) and radiative forcing from aerosol-cloud interactions (ERFaci) (Boucher et al., 2013). Aerosol particle size is an important aerosol physical property. As fine mode aerosols are more related to anthropogenic aerosols (Kleidman et al., 2005; Yan et al., 2017a) when compared to coarse mode ones, fine mode fraction (FMF) is used to distinguish between anthropogenic and natural aerosol types (Remer et al., 2005; Yan et al., 2017a). We define FMF as portion of the fine mode AOD in the total AOD. Yan et al. (2017b) found that particles with a diameter smaller than 2.5 μm ($PM_{2.5}$) have stronger correlation with fine mode AOD (fAOD) than with total AOD (i.e., combined effects of both fine and coarse particles).

AOD is generally obtained by ground-based and satellite-based sensors. The Aerosol Robotic Network (AERONET) is a global network of ground-based remote sensing observations to measure aerosol optical, microphysical and radiative properties (Dubovik and King, 2000; Holben et al., 1998). It uses CIMEL ground-based Sun photometer, a passive remote sensing technique, to measure collimated sunlight at visible and infrared wavelengths, which are then used to calculate total optical depth based on the Beer-Lambert-Bouguer law (Giles et al., 2019; Holben et al., 1998). The uncertainty of AOD has been reported to be 0.01 to 0.02 (Giles et al., 2019). AERONET provides a long-term, continuous database of aerosol-related variables including AOD in total and FMF. It has been widely used as high-quality ground truth data (Choi et al., 2018; Yan et al., 2017a).

Satellite-based AOD has been retrieved from various satellite sensors such as the MODerate resolution Imaging Spectroradiometer (MODIS) onboard Terra and Aqua satellite and the Geostationary Ocean Color Imager (GOCI). Satellite-based aerosol retrieval algorithms are basically based on the inversion algorithm using Look-Up Tables (LUTs) (Levy et al., 2009). LUTs are precomputed with a radiative transfer model given aerosol and environmental conditions. Satellite-observed spectral reflectance data are compared to those of LUTs to find the best fit (least-squares) to retrieve the corresponding aerosol properties.

The existing physics-based AOD retrievals still come with uncertainties inherent in their assumptions. Major issues are mainly from instrument calibration errors, cloud masking errors, assumptions on surface reflectance, and aerosol fine or coarse model selection (Kittaka et al., 2011; Levy et al., 2010; Remer et al., 2005). It was reported that AOD retrievals have large uncertainty over land due to the effect of the reflectance from various land surface materials (Kittaka et al., 2011; Levy et al., 2010; Yan et al., 2017a). Moreover, there is still room for developing a model that is simple and rather easy to continuously update when compared to relatively heavy, complex methods based on radiative transfer models and LUTs. Chen et al. (2020) have jointly retrieved both FMF and AOD using MODIS data using artificial neural networks (ANN). They have reported that their ANN-based models have performed better compared to MODIS products, especially for FMF.

Retrievals of AOD and FMF can be difficult due to the variation of aerosols in time and space. The distribution and variability of aerosols are affected by meteorological conditions. Aerosol loading is governed not only by local aerosol emissions but also by meteorological factors, topographic characteristics, and long-distance transport of aerosols. Mu and Liao (2014) found that meteorological factors can have larger influences on aerosol variations than aerosol emissions. Therefore, it is needed to consider meteorological and geographical factors to quantify aerosols in large, varied study areas. Well-known meteorological factors that affect AOD include relative humidity (RH), temperature, and wind speed (WS). Increasing RH brings increases in aerosol scattering by promoting hygroscopic growth of aerosol particles (Tariq et al., 2021). While low RH (50–80%) does not affect AOD much, very high RH (98–99%) can change about 25% or more of AOD values (Khoshsima et al. 2014). An increase in temperature can affect AOD by enhanced atmospheric convection (Tariq et al., 2021). WS is a factor that can indirectly consider the transport of aerosols from emission sources, but it is still difficult to isolate its effect on AOD (Engström and Ekman, 2010).

In this study, we proposed machine learning-based retrievals of AOD and FMF from GOCI geostationary satellite data using ground-based observations from AERONET and Sun-sky radiometer Observation NETwork (SONET) as reference data. The proposed approach uses meteorological and geographic information related to the temporal cycles of the target variables, in addition to satellite spectral information, and feeds them directly into the model to overcome uncertainties of existing satellite-based aerosol retrieval algorithms, which has not been

Table 1
Input features used in AOD and FMF retrieval models in the first stage (i.e., before feature selection). Spatial resolution means the spatial resolution of the original data. All data were resampled to 6 km spatial resolution. Six VIS channels are at 412, 443, 490, 555, 660, and 680 nm and two NIR channels at 745 and 865 nm. The abbreviations are shown in brackets after the full names.

Source	Variables (abbreviation)	Spatial resolution
GOCI L1	Six VIS and two NIR channels (ch01-ch08)	500 m
	Differences between the reflectance of target day and its minimum value over the past 30 days for each six channels (ch01diff-ch08diff)	
	Normalized Difference Vegetation Index (NDVI)	
	Land sea mask	
GOCI L2	Satellite Azimuth Angle (SAA)	500 m
	Satellite Zenith Angle (SZA)	
	SOLar Azimuth angle (SOLA)	
	SOLar Zenith angle (SOLZ)	
	Relative Azimuth Angle (RAA)	
JAXA AWD30	Terrain elevation from Digital Surface Model (DSM)	30 m
RDAPS	Latent Heat flux (LH)	12 km
	Planetary Boundary Layer Height (PBLH)	
	Relative Humidity (RH)	
	Air Temperature (Temp)	
	Visibility	
	U-wind	
	V-wind	
	Wind Speed (WS)	
GPM	Accumulated Precipitation for 24 h (AP24h)	0.1°
Time-related data	Sine transformed Day Of the Year (DOY)	–

tried in the past studies that retrieved AOD or FMF from satellite data based on machine learning algorithms (Huttunen et al., 2016; She et al., 2020; Liang et al., 2021; Yeom et al., 2021). We aimed at improving model performance presented in the spatial and temporal patterns of AOD and FMF and at investigating machine learning models with detailed feature analysis to provide some explanation about the interactions of input variables. The machine learning models were evaluated and compared with existing satellite-based products on both spatial and temporal domains.

2. Data

2.1. Study area

The study area is the spatial coverage of the GOCI satellite sensor, which is centered on Korean Peninsula and covers the eastern part of China and Japan (22°N–49°N, 112°E–145°E) (Jang et al., 2017). Fig. 1 shows the study area with ground-based observation sites of AERONET and SONET. These areas are well known for serious air quality problems.

2.2. Ground-based observation data

Ground-based observation data from AERONET and SONET were used as reference data. Both AOD and FMF data were taken from ground-based observation stations of AERONET Version 3 Direct Sun Algorithm product (https://aeronet.gsfc.nasa.gov/cgi-bin/webtool_aod_v3) (Giles et al., 2019), which provides observation data in three data quality levels of Level 1.0 (unscreened), Level 1.5 (cloud-screened and quality controlled), and Level 2.0 (quality-assured). In this study, the Level 2.0 aerosol products were used. Likewise, SONET provides AOD and FMF data over China (Li et al., 2018). It guarantees data quality as the same level of the AERONET (Huang et al., 2020). SONET data are beneficial to obtaining more data in China, as AERONET stations are mainly located in Korea and Japan. SONET provides data in four levels-Level 1, 1.5, 1.6, and 2. Since Levels 1.6 and 2 data are not routinely available, the cloud-screened Level 1.5 data were used (Zhang et al., 2019). In this study, a total of 33 AERONET and SONET stations were used from 1 March 2016 to 28 February 2017 to cover all seasons, including 9 additional observations during the Korea-United States Air Quality (KORUS-AQ) field campaign period (from May 2016 to June 2016). The number of ground-based observation sites located on islands or the coast is 3, and the remaining 30 sites are located inland.

2.3. Geostationary ocean color imager (GOCI) data

Table 1 summarizes input features used in machine learning models. Six visible channels (centered at 412, 443, 490, 555, 660, and 680 nm) and two near-infrared channels (centered at 745 and 865 nm) from the GOCI sensor onboard the GEOstationary KOrea Multi Purpose SATellite 1 (GEO-KOMPSAT-1), which is the Korea's first geostationary communications, ocean, and meteorological satellite, were used (Jang et al., 2017; Ryu et al., 2012). GOCI data were obtained from the Korea Ocean Satellite Center (KOSC) (http://kosc.kiost.ac.kr). Level-2 Rayleigh corrected reflectance data for each GOCI band were extracted from GOCI level-1 radiance data through the GOCI Data Processing System (GDPS; http://kosc.kiost.ac.kr/index.nm?menuCd=54&lang=en). As the reflectance depends on satellite and solar viewing angles, sun geometry and satellite viewing geometry information including solar zenith angle (SOLZ), solar azimuth angle (SOLA) and relative azimuth angle (RAA) were also used as input variables. Land-sea mask information was also used from the GOCI data to consider different surface reflectance characteristics between the land and ocean surfaces.

As satellite-observed reflectance is composed of contributions from the atmosphere and the surface, the minimum reflectance technique has been used to determine surface reflectance (Choi et al., 2018; Choi et al., 2016), where surface reflectance is determined as the minimum value of the composited reflectances over a certain time period. In this study, the difference between the observed reflectance and its minimum value over the past 30 days for each channel was used as an input feature regarded as aerosol reflectance. In addition to this, Normalized Difference Vegetation Index (NDVI) was used to give additional information on arid or vegetation areas, which could minimize systematic bias at low AOD especially due to brighter surfaces (Choi et al., 2016; Levy et al., 2013).

2.4. Meteorological variables

Various meteorological features were obtained from the Regional Data Assimilation and Prediction System (RDAPS) data (https://data.kma.go.kr/) to consider meteorological influences on aerosol changes. RDAPS is a numerical weather prediction model, developed by the Korea Meteorological Administration. The RDAPS data source is the boundary fields from Global Data Assimilation and Prediction System (Kang et al. 2021; Park et al. 2020). It provides 3 hourly meteorological forecast fields (i.e., four times a day at 00, 06, 12, and 18 UTC) at a spatial resolution of 12 km and 70 layers up to 80 km. Here, 3-hour averaged latent heat flux (LH), planetary boundary layer height (PBLH), 2 m relative humidity (RH), 2 m air temperature (Temp), 2 m visibility, 10 m U-wind, 10 m V-wind and wind speed (WS) were used in this study. To account for aerosol wet deposition by scavenging, total precipitation was used from the Global Precipitation Measurement (GPM) data (Lee et al., 2011; Textor et al., 2007). Level-3 hourly GPM precipitation data with spatial resolution of 0.1 × 0.1°, named GPM_3IMERGDF, were used. GPM is a joint mission between the Japan Aerospace Exploration Agency (JAXA) and the National Aeronautics and Space Administration (NASA) to observe Earth's global precipitation. It was designed to utilize infrared channel-based estimates in geosynchronous-Earth orbit together with as many low Earth orbiting satellites as possible to compensate for the limitations of using a single satellite data. The precipitation data were obtained from the NASA Goddard Earth Sciences Data and Information Services Center (https://disc.gsfc.nasa.gov/datasets/).

Table 2
The selected input features for each model for AOD and FMF retrievals. b_r means the blue/red channel ratio.

Target	Selected features for LightGBM	Selected features for RF
AOD	ch01diff, ch03diff, ch04diff, ch05diff, ch06diff, ch07diff, ch08diff, ch01, ch02, ch03, ch04, ch05, ch06, ch07, ch08, NDVI, b_r, SOLA, SOLZ, RAA, AP24h, LH, PBLH, RH, Temp, Visibility, U-wind, V-wind, WS, DSM, DOY	ch01diff, ch02diff, ch03diff, ch04diff, ch08diff, ch01, ch02, ch03, ch04, ch06, ch07, NDVI, b_r, SOLA, SOLZ, RAA, LH, PBLH, RH, Temp, Visibility, U-wind, V-wind, WS, DSM, DOY
FMF	ch01diff, ch03diff, ch04diff, ch06diff, ch07diff, ch01, ch03, ch04, ch07, ch08, NDVI, b_r, SOLA, SOLZ, RAA, AP24h, LH, PBLH, RH, Temp, Visibility, U-wind, V-wind, WS, DSM, DOY	ch01diff, ch03diff, ch04diff, ch05diff, ch06diff, ch07diff, ch01, ch04, ch06, ch08, NDVI, b_r, SOLA, SOLZ, AP24h, LH, PBLH, RH, Temp, Visibility, U-wind, V-wind, WS, DSM, DOY

2.5. Other auxiliary variables

Surface elevation data from the JAXA ALOS World 3D-30 m (AW3D30) Digital Surface Model (DSM) were used, as terrain and underlying surface characteristics have been found to be important in aerosol patterns (Cheng et al., 2019). High AOD was often observed in relatively flat, low-lying regions surrounded by elevated terrain due to weak dispersion of aerosols (Cheng et al., 2019; Shi et al., 2018). In addition, Day Of the Year (DOY) was used to incorporate temporal dependence of AOD into the model. In East Asia, it is crucial to consider seasonality in air quality studies (Kang et al., 2021; Park et al., 2020).

2.6. Data for comparison

AOD and FMF from the Yonsei aerosol retrieval (YAER) aerosol products at 550 nm with a 6 km × 6 km spatial resolution were used for comparison. The GOCI aerosol products are derived by the GOCI YAER version 2 algorithm (Choi et al., 2018). The GOCI YAER products are calculated based on the radiative transfer model. The discrete ordinate radiative transfer code of the libRadtran software package models the top of the atmosphere (TOA) reflectances from AOD. The aerosol products are then calculated through the inversion process from TOA reflectances based on LUTs (Choi et al., 2016). YAER aerosol products are produced only when the number of valid pixels is more than 28 in the process of aggregation from 500 m to 6 km (12 × 12 pixels). After that, TOA reflectance and its standard deviation within 6 km were examined in the next masking step so that no results were produced for pixels with bright and heterogeneous surfaces (Choi et al., 2016).

MODIS AOD and FMF products were also used for comparison. We used the Level-2 daily swath products with 10 km spatial resolution (MOD04_L2 for Terra and MYD04_L2 for Aqua; Collection 6.1). They are produced based on two MODIS aerosol algorithms, which are so-called the Dark Target (DT) retrieval for dark surfaces over land and ocean, and the Deep-Blue (DB) algorithm mainly for bright-desert regions as aerosol signals stand out and bright surfaces are relatively dark in the near-UV band. Details on algorithms can be found in for the DT algorithm (Levy et al., 2009) and for the DB algorithm (Hsu et al., 2004). The quality assurance confidence flag (QAC) values of MODIS AOD data have 3 (Very Good), 2 (Good), 1 (Marginal), and 0 (bad quality) (Levy et al., 2013). High quality data were used for both MODIS DT (QAC = 3 over land and QAC ≥ 1 over ocean) and DB (QAC of 2 and 3).

3. Methods

3.1. Data preprocessing

Multiple data preprocessing steps were applied to construct training data that were fed into data-driven machine learning models. The steps involve variable transformation, derivation of new variables, and collocation at different temporal and spatial resolutions. For a proper comparison, it is needed to adjust AOD values to a wavelength of each sensor. The AERONET AOD measured at 550 nm were corrected by using the following power law equation (Bibi et al., 2015):

$$AOD_c = AOD_a \left(\frac{c}{a}\right)^{-\alpha} \quad (1)$$

where AOD_c is the corrected AOD, AOD_a is the AERONET AOD, c is a common wavelength of 550 nm, a is 500 nm of AERONET, and α is Angstrom exponent of 440–870 nm. Next, the corrected AOD and FMF at 550 nm were collocated with GOCI data. Since GOCI observes the Earth at hourly intervals from 00:30 to 07:30 UTC (09:30–16:30 Korea Standard Time) (Choi et al., 2018), AERONET observations within 30 min before and after the GOCI observation time were averaged.

The hourly meteorological fields were acquired by linear interpolation from analysis fields with 6-hour intervals, and then they were resampled to a 6 km spatial resolution through spatial bilinear interpolation to match up with the GOCI grid. The GPM data were used as the accumulated precipitation over the past 24 h before the satellite observations (termed AP24h) for every hour. The JAXA AW3D30 DSM was aggregated to a 6 km resolution to match the GOCI grid from a 30 m resolution. DOY was transformed to have a value between −1 and 1 through sine transformation, which represent a high peak (value 1) is a warm season while a low peak (value −1) means a cold season.

3.2. Cloud masking

AOD and FMF retrievals were performed for clear sky conditions. The cloud mask was derived from the YAER AOD product (Choi et al., 2018). There are multiple steps for masking clouds and other contaminated pixels over the ocean and land. As a first step, pixels are classified as cloud if sequential, multiple conditions are met, which are applied to GOCI reflectance images of 0.5 × 0.5 km resolution. The conditions are basically to measure variations in a certain window: high variability in the window is determined as clouds on land and vice versa. As a next step, if the number of available pixels in a 12 × 12 pixel window is larger than 72, the darkest 20% and brightest 40% of reflectances at 490 nm are removed, and the remaining pixels are averaged and aggregated to 6 × 6 km resolution. Then, additional masking at 6 × 6 km resolution is applied for detailed classification. Detailed descriptions of cloud masking and the aggregation processes are found in Choi et al. (2018).

3.3. Feature selection

The recursive feature elimination (RFE) method was applied to select input parameters that are useful to predict AOD and FMF. RFE first computes the importance of each feature for the initial set of variables, then removes a variable with the least importance, and recalculate the importance for the pruned set. This process is repeated to obtain the optimal set of input features. Each model with the default parameter setting was used for this task, and the model at each iteration was evaluated based on cross-validation. Feature importance was examined using the number of times each variable is used during training for the LightGBM model, and mean decrease in prediction error (mean squared error) for the RF model. The models were trained with a subset of input features selected by the feature selection method (Table 2).

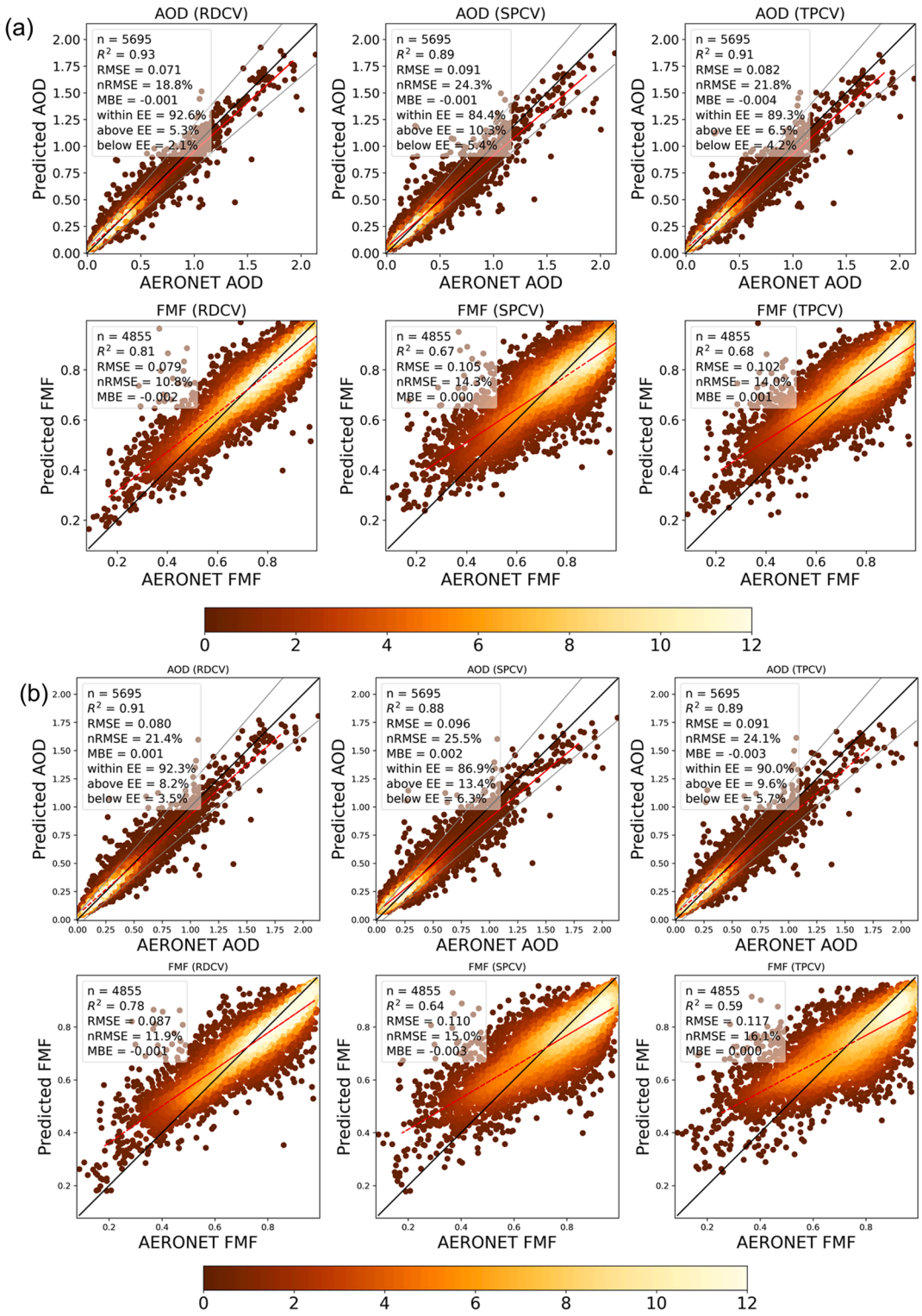

Fig. 2. Scatter plots of *in-situ* observations (x-axis) and prediction (y-axis) from (a) LightGBM and (b) RF for each cross-validation for AOD (top) and FMF (bottom). The red solid line is a line of the best fit to the scatter plot, and the black dotted line is an identity line. The gray lines represent the expected error of MODIS DT AOD. The dot color means the point density from dark red (low) to white (high), estimated by a Gaussian kernel density estimation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Machine learning models

Two machine learning techniques were applied to predict AOD and FMF, which are Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) regression models. LightGBM is a gradient boosting-based decision tree ensemble algorithm, which basically builds decision trees by dividing training data in a direction of decreasing the gradient of the loss or cost function (i.e., reducing errors), and then combines the decision trees (i.e., ensemble) to obtain final prediction. LightGBM works differently in growing trees and is more efficient in training time than other existing boosting-based algorithms (Ke et al., 2017; Pham et al., 2021). Existing gradient boosting-based decision tree algorithms

Table 3
Accuracy assessment results of each model for AOD for the separate test data.

Accuracy measure	LightGBM	RF	YAER	MODIS AOD (DT)	MODIS AOD (DB)
Number of samples	613	613	613	104	136
R^2	0.93	0.92	0.89	0.82	0.69
RMSE	0.079	0.086	0.130	0.186	0.167
nRMSE	18.9 %	20.7 %	30.9 %	38.2 %	42.6 %
MBE	−0.001	−0.003	−0.081	0.101	−0.033
% within EE	89.1 %	86.3 %	52.2 %	49 %	48.5 %
% above EE	8.0 %	10.3 %	4.6 %	44.2 %	22.1 %
% below EE	2.9 %	3.4 %	43.2 %	6.7 %	29.4 %

Table 4
Accuracy assessment results of each model for FMF for the separate test data.

Accuracy measure	LightGBM	RF	YAER	MODIS FMF (DT)
Number of samples	259	259	259	67
R^2	0.78	0.74	0.44	0.18
RMSE	0.098	0.110	0.309	0.456
nRMSE	13.9 %	15.0 %	43.8 %	67.3 %
MBE	0.030	0.013	−0.255	−0.246

use level (depth)-wise tree growth. They build the trees' depth first, which often takes a long time to optimize the trees. On the other hand, LightGBM grows trees in a leaf-wise manner with the Gradient-based One-Side Sampling (GOSS) sampling method. GOSS focuses on data with large gradients, while data with small gradients are randomly dropped and weighted with a constant value. As the approach searches for a smaller portion of samples rather than the whole data, it shows high efficiency, often resulting in comparable or better accuracy than others (Ke et al., 2017). In addition, LightGBM is trained in a way that avoids the local minimum problem by using a subsample of the training data to find the global optimum. In this study, LightGBM was implemented with a Python package, lightgbm, in a Python 3 scikit-learn environment. Hyperparameters were optimized using a grid search method based on 5-fold cross-validation. The grid search method tests various combinations of the hyperparameters to find the optimal hyperparameters based on the model performance (here R^2 value). The model hyperparameters tuned using the grid search method include max_depth, min_data_in_leaf, n_estimators, and num_leaves. The max_depth parameter determines the maximum depth of the tree model. The min_data_in_leaf parameter is the minimum number of samples required for a leaf, which is used to control overfitting. It is affected by the number of training samples and the num_leaves parameter. While setting this value high can avoid growing too deep trees, it may cause underfitting. Thus, it needs to be optimized by hyperparameter tuning. The n_estimators parameter is the number of trees, which is related to the learning_rate parameter. The learning rate typically increases with the decreasing number of trees. The num_leaves parameter indicates the number of leaves, which is a major parameter that determines the complexity of the tree model.

RF builds a number of decision trees on randomly drawn subsets of training data with replacement, which is the Bootstrap Aggregation (or bagging in short), and aggregates the outputs of the trees (Breiman, 2001; Gumma et al., 2020). Trees use the subsets of given features and samples, making splits along decision nodes by increasing the homogeneity of sub-nodes based on a criterion to measure the purity of the split. The best split at each node of a tree during training is searched with a random subset of input features or all features. The randomness for both training samples and input features introduced in the forest is intended to reduce the variance of the forest as individual decision trees tend to be diverse and overfit samples. The final prediction for regression is made by the ensemble of the trees by averaging predicted values of the trees. RF was implemented with the scikit-learn module in a Python 3.8.5 environment (Pedregosa et al., 2011). Model parameters tuned were the number of trees (i.e., n_estimators), the size of the random subsets of input features to consider at each node (i.e., max_features), the maximum depth of the tree (i.e., max_depth), and the maximum number of leaves (i.e., max_leaf_nodes). The optimal parameters were set as: n_estimators = 700, max_features = the square root of the number of features, max_depth = 12 and max_leaf_nodes = 600 for AOD and n_estimators = 600, max_features = the number of input features, max_depth = None and max_leaf_nodes = 300 for FMF. If the max_depth is set as "None", then nodes are expanded until all leaves are pure. Mean squared error was used to measure the quality of the split.

3.5. Evaluation methods

Accuracy measures including R^2, Root Mean Squared Error (RMSE), Normalized RMSE (nRMSE), and Mean Bias Error (MBE) were used to evaluate the models developed in this study. Bias is calculated as the difference between the mean of the difference between predicted and observed values. Additionally, for AOD, Expected Error (EE) was also used. EE refers to the expected error of AOD that is determined by the solar zenith angle and satellite zenith angle (Choi et al., 2018). We used the following EE from MODIS DT AOD algorithm for comparison (Levy et al., 2013).

$$EE_{MODISDT} = \pm(0.15 \times AERONETAOD + 0.05) \quad (2)$$

A total of 6308 data samples for AOD and 5114 for FMF were obtained. For model validation, we performed random, spatial, and temporal cross-validation to show the spatial and temporal transferability of the approach. Each validation is named random 10-fold cross-validation (RDCV), spatial 10-fold cross-validation (SPCV) and temporal 10-fold cross-validation (TVCV). The adopted three cross validation approaches have been used in recent studies to show the robustness of their proposed approaches including their transferability (i.e., generalization) (Kang et al., 2021; Wang et al., 2021, Huang et al., 2018; Wei et al., 2021; Reitz et al., 2021). The 10-fold CV divides the whole data into 10 subsets, where 9 subsets are used for training and the remaining 1 subset for validation. Then, the validation result for the whole dataset is obtained by averaging the 10 results of the 10 folds. The temporal cross validation randomly divides the data by date, which means separate data by dates for each fold. The spatial cross validation is done by sites in the same manner. Additionally, the model was analyzed using separate test sites that are not used in training. At least one station was extracted from South Korea, China, and Japan for separate test sites, so samples from three AERONET and one SONET stations were used for model analysis.

3.6. Model interpretation methods

Tree-based machine learning methods allow us to interpret the model responses to input features, which helps to understand the model's decision processes between input features and the target. The model interpretation methods used in this study are feature importance, partial dependence plots, and feature interaction values based on SHapley Additive exPlanation (SHAP) regression values (Lundberg and Lee, 2017). The SHAP method uses Shapley values from coalitional game theory, which is to represent the contribution of each feature to the model prediction of each data instance. SHAP feature importance is basically a measure of feature importance based on the magnitude of absolute Shapley values of input features. SHAP dependence plots give a global interpretation on the effect of an input feature on the model predictions. They are obtained by isolating the influence of one input variable, while all the other variables are fixed, and then we get the averaged response of the model for each input feature to analyze. They show the averaged model response to each feature. While partial dependence plots show the global averaged effect of each input feature to the model predictions, SHAP regression values can quantify each

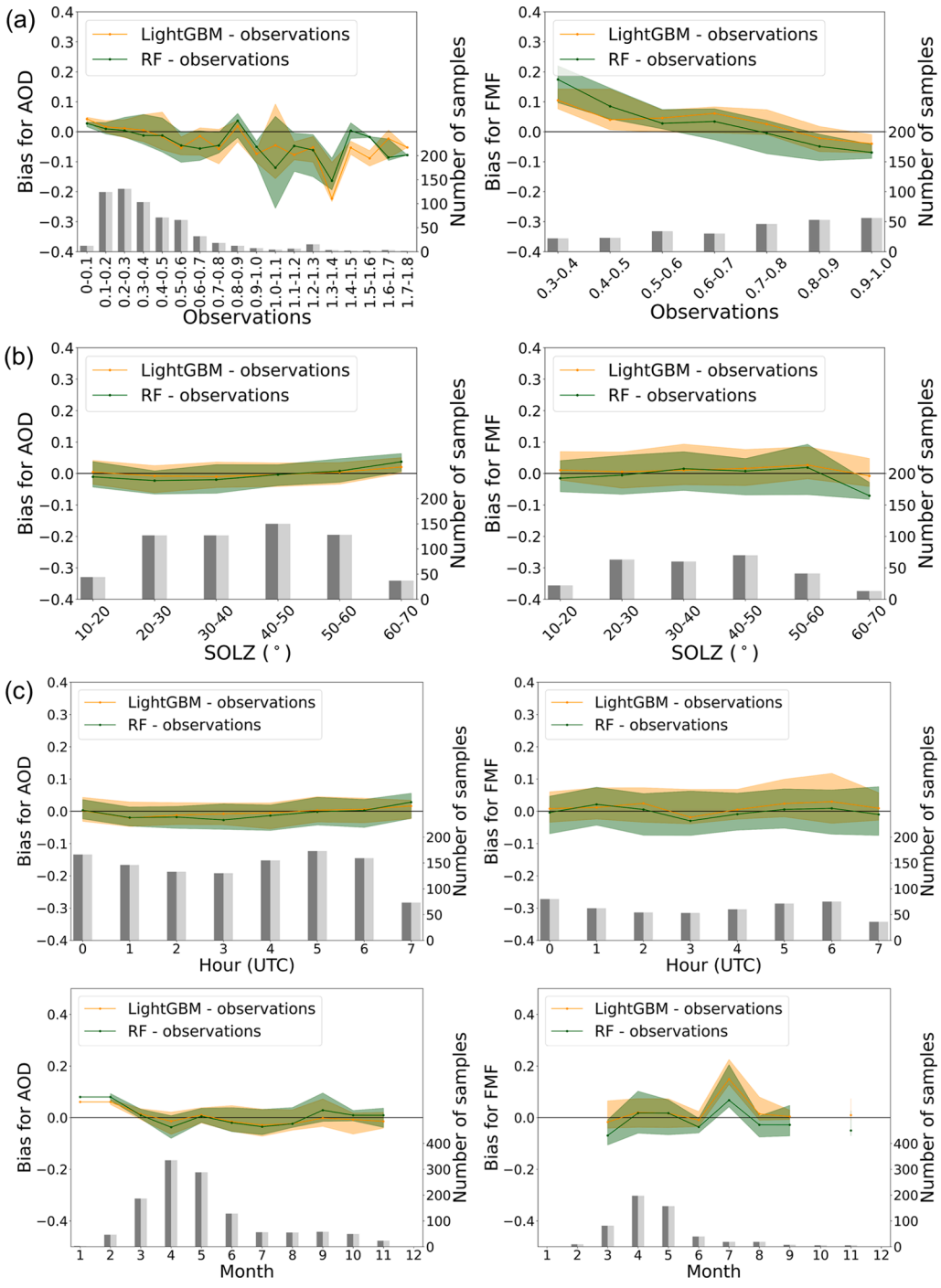

Fig. 3. Bias distribution between AOD (left column) and FMF (right column) of the LightGBM and RF models compared to the observations of AERONET and SONET for independent test data by (a) observation data, (b) solar zenith angle (SOLZ), and (c) hour (upper panel) and month (lower panel). Colored solid lines are the median of the bias with the interquartile range as shaded area. Gray bars represent the number of samples for LightGBM (left bar of dark gray) and RF (right bar of light gray).

input features contribution to every single model outcome by accounting for the feature interaction effects (Lundberg and Lee, 2017). Therefore, SHAP values can provide a deeper understanding of model decisions.

4. Results and discussions

4.1. Model performance

4.1.1. Cross-validation results

Fig. 2 shows the results of three types of cross-validation (i.e., RDCV, SPCV, and TPCV) for the LightGBM and RF models. In the upper panel of Fig. 2a, the LightGBM model explained 93, 89, and 91% of the variations in AERONET AOD values for RDCV, SPCV, and TPCV, respectively, indicating no significant systematic bias observed between the validation results. These results were better than those reported in Levy et al. (2013) with MODIS AOD (0.74 for land and 0.88 for ocean) and in Choi et al. (2018) with GOCI-based YAER AOD (0.83 for land and 0.79 for ocean). In addition, a high fraction of the predicted AOD (92.6, 84.4, and 89.3%) were within the EE envelopes, which were higher than those reported in Levy et al. (2013) with MODIS AOD (69.4 % for land and 76.16 % for ocean) and in Choi et al. (2018) with GOCI-based YAER AOD (60 % for land and 71 % for ocean). However, it should be noted that the accuracies cannot be directly compared to the literature as the data for evaluation were different by study. For FMF, the models performed worse than AOD, resulting in R^2 of 0.81 (RDCV), 0.67 (SPCV),

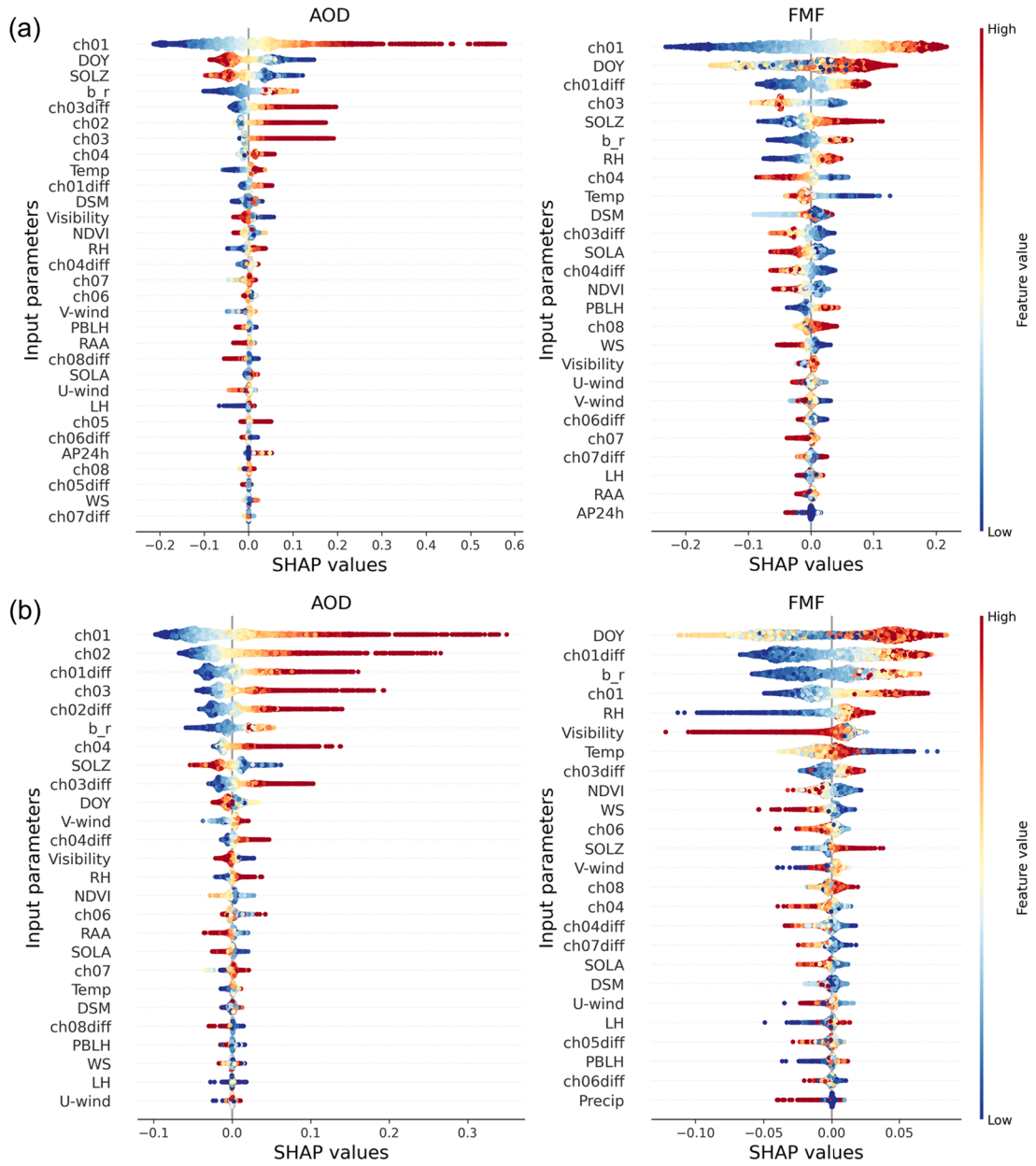

Fig. 4. SHAP feature importance combined with feature effects for AOD (left) and FMF (right) for (a) LightGBM and (b) RF models. Each point represents a Shapley value for a feature and an instance. The color shows a high and low degree of feature values. Overlapping points are scattered in y-axis direction. The input parameters are ordered according to their importance. The color means that a value of a given feature is high or low. The horizontal distribution of dots indicates that a given feature contributed to higher or lower predictions along the x-axis.

and 0.68 (TPCV), as shown in the lower panel of Fig. 2a. These were better than those from Kleidman et al. (2005) who compared AERONET to MODIS FMF and reported R^2 of 0.6 and from Choi et al. (2018) with R^2 of around 0.39 for land and 0.25 for ocean. Very low FMF values tended to be overestimated, and they were found to mainly come from Beijing sites. This is likely caused by high reflectance of urban surfaces, which has been a problem in retrieving aerosol products (Gupta et al., 2016). In Fig. 2b, it is shown that the results of the RF model were almost similar to the those of LightGBM for AOD, and slightly worse for FMF, but better than those previously reported (Choi et al., 2018; Levy et al., 2013). Both models tended to have larger error in the range with less data for both AOD and FMF.

Some outlying samples in the validation results were analyzed to investigate the main factors that influenced them. In terms of AOD, model performances were slightly better for RDCV, followed by TPCV and SPCV. In general, the fraction of the above EE (i.e., samples that fall above the upper EE bound) was higher than the below EE (i.e., samples that fall below the lower EE bound), which was even higher in SPCV. It was found that Japan sites account for about 10 % of the entire samples, but many of the samples in the above EE of SPCV came from Japan sites. In other words, as high AOD occurred more often in South Korea and China than in Japan, the models tended to overestimate AOD in Japan, where AOD was relatively low throughout the year. In addition, some AERONET AODs around 0.25 were extremely overestimated by over 1.0 for both models. Such cases occurred at the Baengnyeong island station in South Korea, and the ch01 feature contributed most to the overestimation. Reflectance features were found to be high due to the unmasked thin clouds, which resulted in exceptionally overestimated AOD. In addition, low FMF values were generally overestimated, which was found to be mostly from Beijing sites.

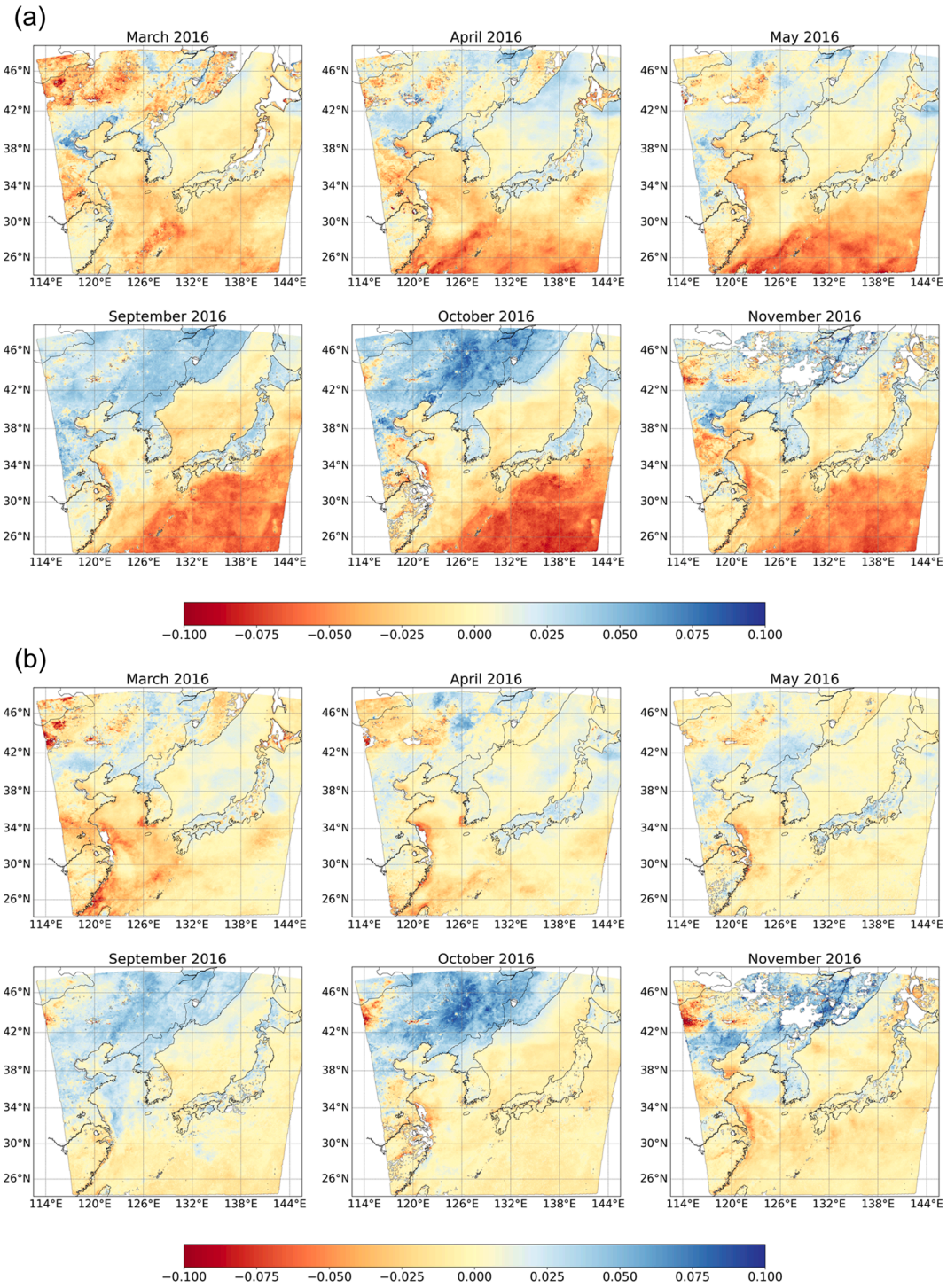

Fig. 5. Spatial distribution of monthly AOD differences between the original model and the test model without the channel difference input features. Each model was generated by (a) LightGBM and (b) RF. The positive value means that the test model has higher AOD than the original model, and vice versa.

4.1.2. Model analysis with separate test sites

The final models were analyzed using the separate test sites, which were not used in training the models. The model results were compared to the corresponding AOD and FMF data from GOCI-based YAER and MODIS, as shown in Tables 3 and 4. It should be noted that the comparison results can be affected by the selection of test stations. For AOD, both the LightGBM and RF models performed slightly better, resulting in R^2 of 0.93 than the other products for YAER, MODIS DT, and DB with R^2 of 0.89, 0.82, and 0.69, respectively (Table 3), but showed a much higher fraction within EE with 89.1 and 86.3 % for LightGBM and RF when compared to 52.2, 49, and 48.5 % for YAER, MODIS DT and DB, respectively. The LightGBM and RF models did not feature a substantial bias with MBE of −0.001 and −0.003, respectively, whereas YAER and MODIS DB showed a relatively greater underestimation of AOD with a large negative MSE (-0.081 and −0.033, respectively) and a much higher fraction of below EE, and MODIS DT showed a much higher overestimation of AOD (MAE = 0.101). It should be noted that the number of samples used for this analysis was smaller for MODIS. In terms of FMF, it was also found that the LightGBM and RF models showed better performance over the others, with much higher R^2 of 0.78

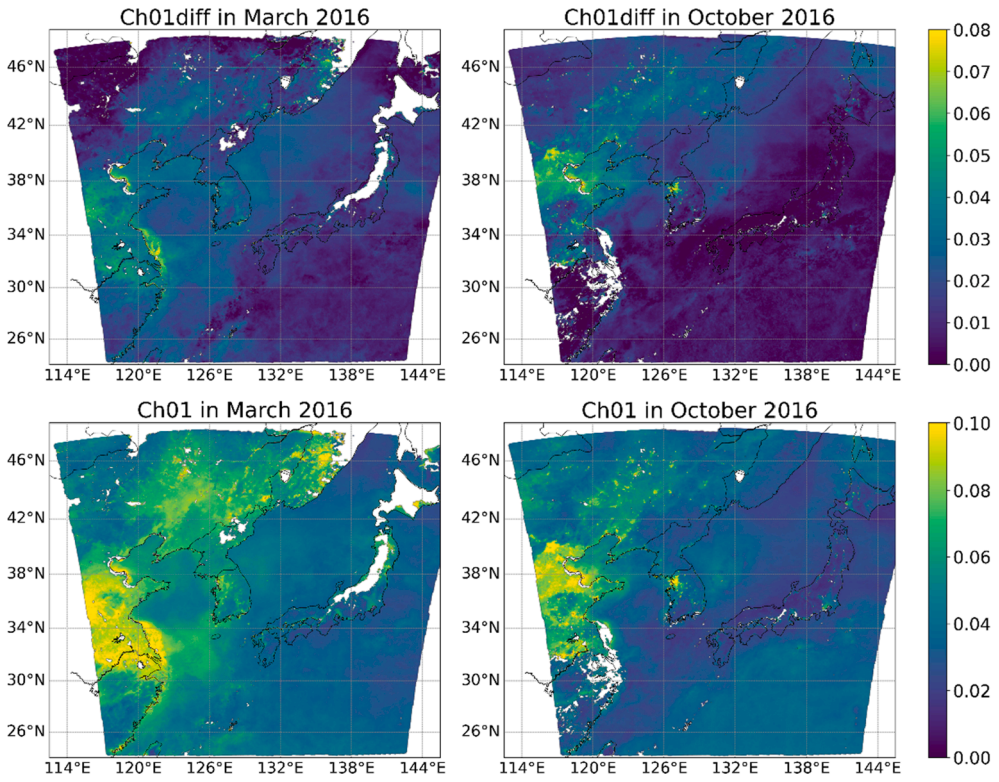

Fig. 6. Spatial distribution of monthly ch01diff (upper panel) and ch01 (bottom panel) input features.

and 0.74 compared to 0.44, and 0.18 for YAER and MODIS DT, respectively (Table 4). Meanwhile, the LightGBM and RF models tended to have a slightly positive bias (MSE: 0.030 and 0.013), while YAER and MODIS have a large negative bias (MSE: −0.255 and −0.246).

Fig. 3 depicts the biases between AOD and FMF of the RF and LightGBM models and observations from AERONET and SONET for the separate test data by observation data, SOLZ, hours, and month. In the left column in Fig. 3a, both models generally had a small positive bias at lower AOD and more negative, varying bias at higher AOD with less samples, which is consistent with the cross-validation results and previous studies (Chen et al., 2020; Choi et al., 2018). It was found that a small positive bias was due to cloud contamination, which was also a major problem identified in the past studies (Choi et al., 2018; Levy et al., 2013). For FMF, both models had a high positive bias at low FMF with less samples (right column in Fig. 3a), which was also observed in the cross-validation results. In Fig. 3b, the bias was small on average and did not vary much with SOLZ. Several studies (i.e., Choi et al., 2018; Levy et al., 2010; Sayer et al., 2013) reported an increased bias at higher scattering angles (i.e., near-backscattering geometry), as the contribution of the atmosphere decreases due to the reduced atmospheric path length, but this was not found in this study, which means that the models were well generalized across various viewing geometry. Fig. 3c shows that there was no significant difference in bias depending on hours for both models, but there was an abrupt increased bias in July for FMF. It was found that the cases in July were small and mostly from Japan sites, and one of them with very low aerosol loading was significantly overestimated by the models, which led to a sharp increase in bias.

4.2. Model interpretation

Fig. 4 shows SHAP-based feature importance results for AOD and FMF for the LightGBM and RF models, respectively. Each dot is the contribution of each instance of a given feature to each model output. For AOD, the ch01 feature had dominant importance over all the other features especially for LightGBM. At the 412 nm band (i.e., ch01), aerosol signals are brighter and better discernible from the surface, even bright regions (i.e., visually bright, and high surface reflectance at visible channels), when compared to longer wavelengths (Choi et al., 2016; Kaufman et al., 1997; Levy et al., 2013). It is thus seen that high values of ch01 caused higher predictions, whereas low ch01 caused lower predictions. Meanwhile, the contributions of features are not highly biased in the RF model. Meanwhile, RH showed a high contribution to the AOD retrieval for both models. The higher the RH, the higher the predicted AOD (Fig. 4a). For FMF, model predictions were simultaneously influenced by multiple features in the LightGBM model. It was found that DOY, ch01, ch01diff and b_r mainly contributed to FMF retrieval. As fine and coarse mode AODs respond differently to different wavelengths (Che et al. 2015; Fotiadi et al. 2006; Mai et al. 2018), it seems that the b_r (ratio of shorter wavelength (blue) and longer wavelength (red)) have jointly contributed to estimating FMF. In terms of meteorological factors, the contribution of RH was relatively high, which is related to enhanced aerosol scattering due to hygroscopic growth of aerosol particles with increasing RH (Ng et al., 2017; Yoon and Kim, 2006). Temperature was not the most important feature globally, but with high SHAP values, it was also an important feature for a certain range of FMF values.

It was found that DOY was an important feature in retrieving both AOD and FMF. Although only one year data were used in this study, it should be noted that aerosols in the study area has a strong seasonality (i.e., high AOD in winter and spring), and thus it is not surprising to have DOY as an important feature (Kang et al., 2021; Park et al., 2020). In addition, DOY interacts with other meteorological variables and together contributed to the retrieval of AOD and FMF (Fig. S1). It was also confirmed that, without DOY, the model performance was not significantly different from the original model in each cross-validation (Tables S1 and S2).

To further investigate the role of the channel difference features (i.e., ch01diff-ch08diff) in considering the regional surface characteristics, test models were constructed by excluding the channel difference features and compared with the original models. The models followed the

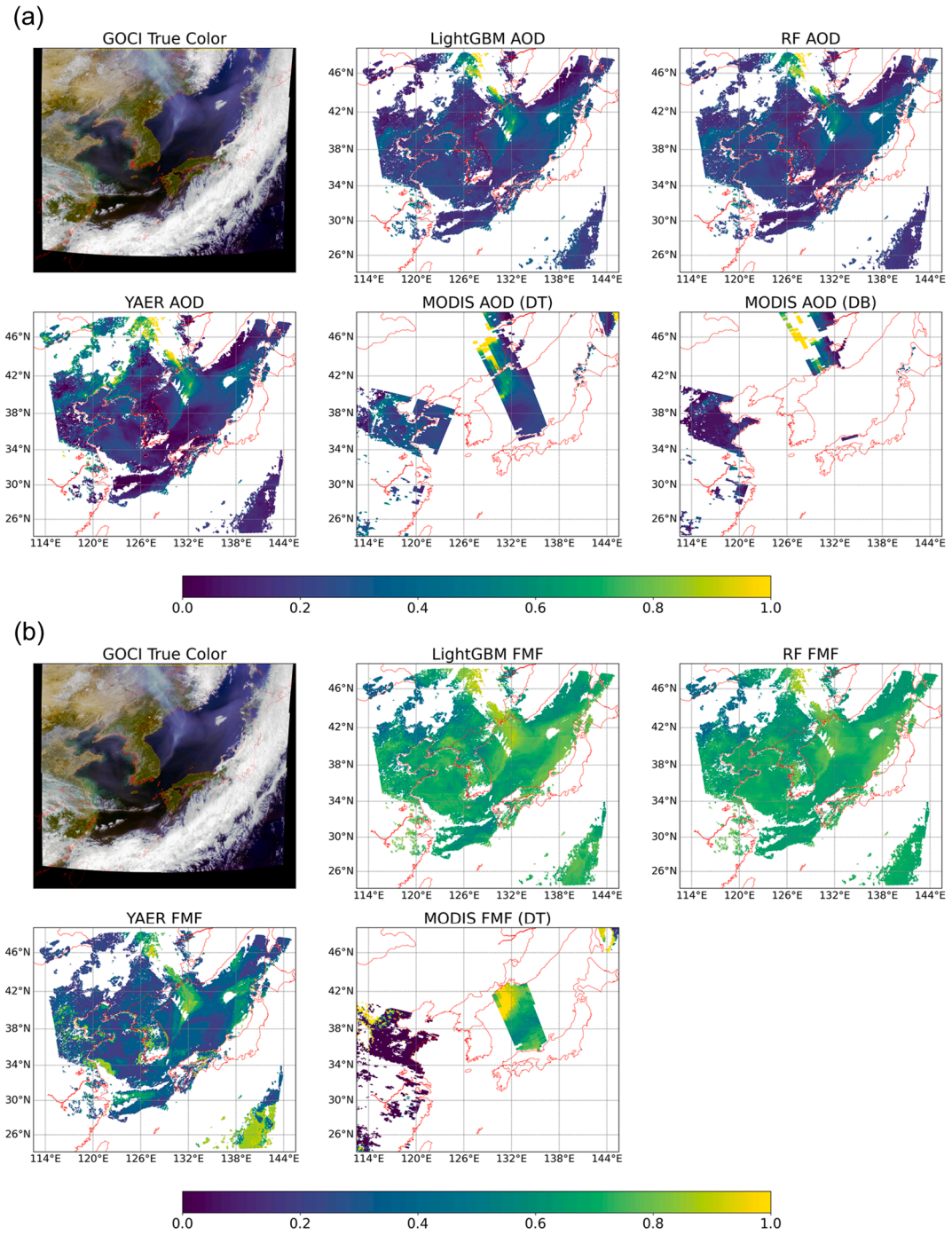

Fig. 7. Map comparison of (a) AOD and (b) FMF from LightGBM, YAER, MODIS DT and DB algorithms on 17 May 2016 at 4 UTC with GOCI True Color imagery.

same approach described in Section 3.3. Fig. 5 shows monthly AOD differences between the original and the test model results for LightGBM and RF, respectively. It was observed that there are spatial differences over land in the estimation of AOD according to the season. In the spring season (from March to May in the upper panels in Fig. 5a and Fig. 5b), the test model tended to estimate AOD higher than the original model in the upper left region in the study area, where desert and semiarid regions exist. It was found to be due to the absence of the channel difference features that contributed to decreasing AOD over the bright underlying surfaces. As shown in Fig. 6, as ch01diff was close to 0, which means that there were no or very low aerosol loadings, it contributed to lower AOD in the original model, but this was not in the test model. In autumn (from September to November in the bottom panel in Fig. 5a and b), positive differences were observed in the upper center of the study area, which represents that the test model tended to underestimate AOD compared to the original model. It was also observed that the channel difference features contributed to increasing AOD in this case, but the test model has no information on the channel difference and produced decreased AOD. Another case of the underestimation of the test model was observed in the Beijing region in the left side of the study area. The region showed both high reflectance in ch01 and high ch01diff, which together leaded to higher AOD than the test model with no channel difference features. This suggests that the channel difference features need to be used in such retrievals; otherwise, aerosol parameters may be overestimated by surface influences.

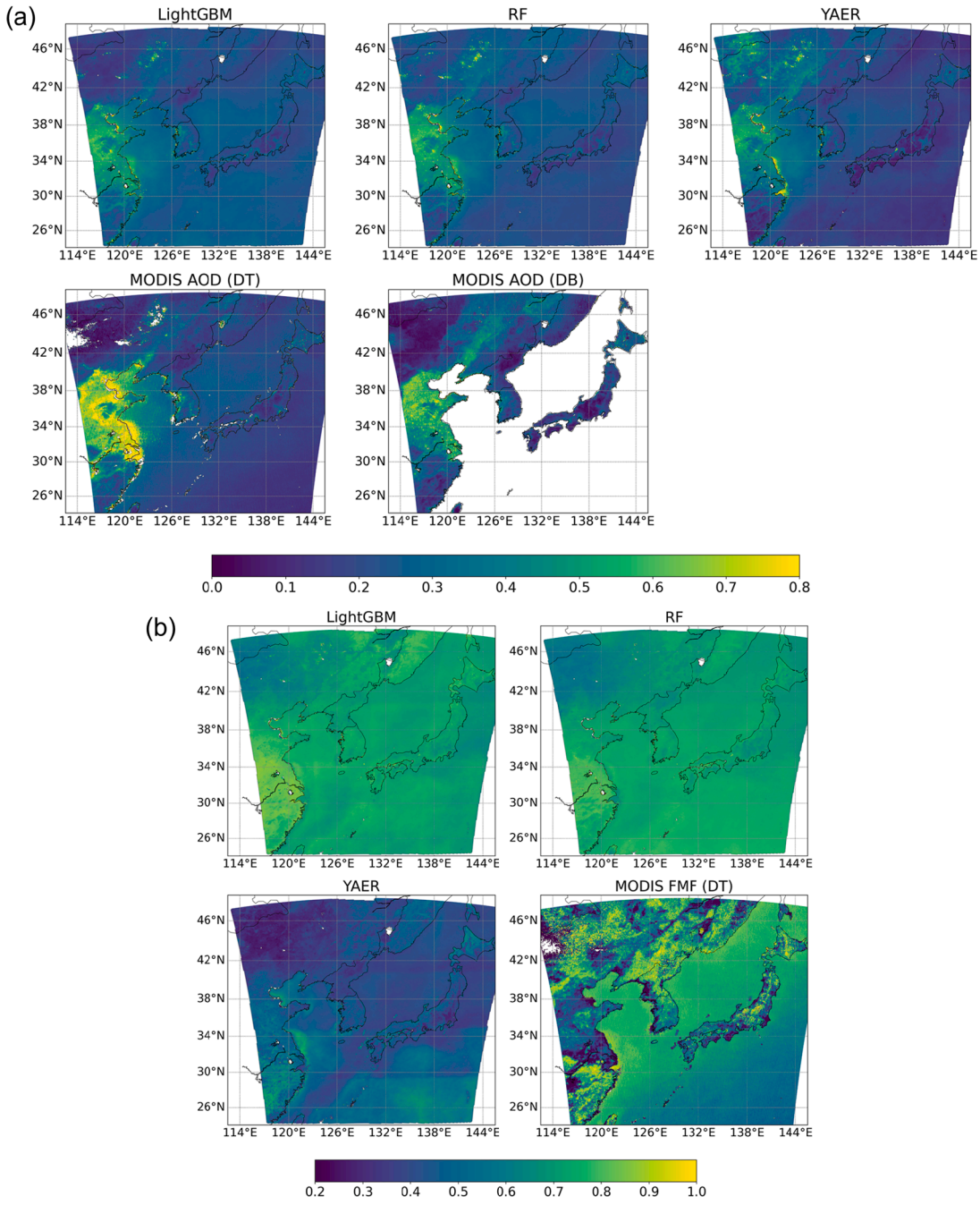

Fig. 8. Averaged spatial patterns of (a) AOD and (b) FMF from LightGBM, RF, YAER, and MODIS DT and DB algorithms for the entire study period (March 2016 – February 2017).

4.3. Spatial distribution with satellite-based products

Fig. 7 shows comparisons of AOD and FMF predicted by the LightGBM and RF models with the corresponding GOCI true color image, GOCI-based YAER, and MODIS DT and DB data on 17 May 2016 at 4 UTC. In this example, a massive smoke aerosol plume is clearly visible in the center top of the true color image. The event was originated from Siberian wildfires, occurred in the Russian forest, which could be transported by strong westerlies (Choi et al., 2019). Smoke from biomass burning has large amounts of fine particles resulting in high values both AOD and FMF (Oros et al. 2006). In Fig. 7a, the aerosol plume was well captured by the proposed models and also YAER and MODIS. It was also evident that in northeast China (i.e., upper left part of the study area), YAER tended to estimate AOD higher compared to the LightGBM and RF models, and MODIS algorithms were limited in the spatial coverage due to its polar-orbiting imaging. It should be also noted that sufficient observations were not available enough over the ocean, which would lead to higher uncertainty over the ocean compared to the land. Nonetheless, the LightGBM and RF models can provide accurate, spatially continuous distribution of aerosol properties at higher temporal resolution, especially than MODIS DB and DT algorithms by using geostationary satellite images, which would be more useful in various applications as aerosols rapidly change in time and space.

Fig. 8 shows the annual mean spatial distribution of AOD and FMF from LightGBM, RF, YAER, and MODIS DT and DB algorithms for the entire study period (March 2016–February 2017). Fig. 9 shows the spatial distribution of R^2 values as model performance to AERONET data. All algorithms showed similar patterns for AOD (Fig. 8a), but it

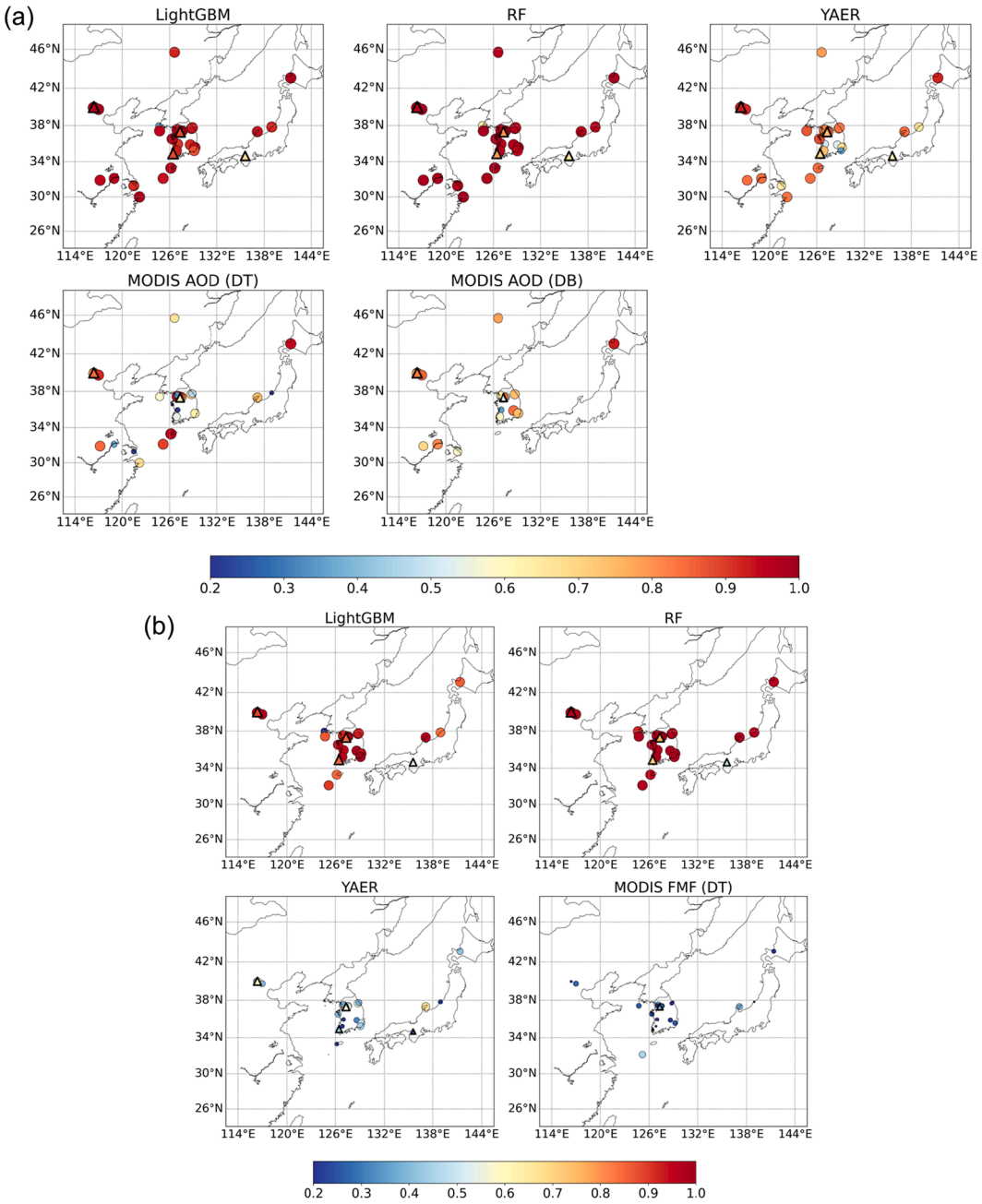

Fig. 9. Spatial patterns of model performance (R^2) for (a) AOD and (b) FMF from LightGBM, RF, YAER, and MODIS DT and DB algorithms for the entire study period (March 2016 – February 2017). Triangles with thick solid line indicate the test sites. As the R^2 increases, the size of the symbols increases and their color changes from blue (low R^2) to red (high R^2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

was also found in the annual mean map that the YAER algorithm showed slightly higher AOD in the upper left region of the study area compared to the other models. In Fig. 9a, LightGBM and RF models were generally well-matched with AERONET data showing high R^2 values, but the YAER and MODIS algorithms showed slightly lower R^2 than the other models for the entire study period (Table 5). It should be noted that the number of samples of MODIS algorithms matched with AERONET data was much smaller due to its polar orbiting imaging, which led to less or no samples available at some sites. Fig. 8b shows the spatial distribution of annual mean for FMF. It was evident that LightGBM and RF models showed higher FMF values than the YAER model in general. Meanwhile, the spatial distribution of MODIS FMF over ocean looks similar with those of LightGBM and RF models, but it showed spatially inhomogeneous patterns. This was reported in past studies that MODIS FMF has much lower accuracy over land than over ocean due to surface contributions (Levy et al., 2013; Chen et al., 2020). In the comparison to the AERONET data (Fig. 9b), the YAER and MODIS algorithms showed overall lower R^2 values compared to the other models. The models proposed in the present study yielded moderate R^2 values at the test stations (Table 5), which were discussed in the validation results in Section 4.1.2.

It should be noted that while a wide range of AOD were well detected by the models, there are clearly limitations of the models for cloud masking. We found that some of the dust area with high AOD have been masked out. This is a factor limiting the model's ability to estimate situations such as urban areas like Beijing, where high AOD frequently occur, as high AOD data are not properly considered in the model. It has been a challenge to distinguish aerosols from clouds and bright surface

Table 5
Model performance (R^2) for the averaged spatial patterns of AOD and FMF from LightGBM, RF, and YAER algorithms for the entire study period (March 2016 – February 2017) to AERONET AOD and FMF data at the test sites. The hyphen indicates no samples available at the site.

Site location		AOD				
Latitude	Longitude	LightGBM	RF	YAER	MODIS DT	MODIS DB
34.651°	135.591°	0.68	0.66	0.62	–	–
37.312°	127.31°	0.88	0.80	0.74	0.62	0.51
34.913°	126.437°	0.89	0.79	0.70	–	–
40.005°	116.379°	0.96	0.96	0.93	0.85	0.83
Site location		FMF				
Lat	Lon	LightGBM	RF	YAER	MODIS DT	
34.651°	135.591°	0.54	0.49	0.26	–	
37.312°	127.31°	0.85	0.76	0.57	0.39	
34.913°	126.437°	0.85	0.71	0.43	0.02	
40.005°	116.379°	0.89	0.97	0.61	–	

(Choi et al., 2018; Levy et al., 2013). For FMF shown in Fig. 7b, the LightGBM and RF models generally had higher FMF than YAER. FMF retrieval is still very challenging to extract from satellite data. Physical models (i.e., satellite products) have still shown very poor performance. Choi et al. (2018) have reported that FMF retrievals have higher bias and lower correlation coefficients than AOD retrievals, which were consistent with the findings of this study. Meanwhile, the LightGBM and RF models yielded the performance similar to the MODIS DT algorithm over the ocean in the near center of the study area, but MODIS DT over land resulted in significantly low FMF. Levy et al. (2010; 2013) have reported that MODIS DT had considerable uncertainties especially over land.

5. Conclusions

This study applied machine learning models to estimate AOD and FMF using spectral channels from GOCI geostationary satellite images, their channel differences, and meteorological and geographical data. The results showed that the models were more consistent with AERONET ground-based observation data than the existing physical model-based products. AOD and FMF predicted by the LightGBM and RF models did not show a significant systematic bias in the spatiotemporal cross-validation, and generally in a good agreement with AERONET, especially for AOD. The findings of this study showed the potential use of machine learning models for reliable retrievals of AOD and FMF from GOCI geostationary satellite data and various meteorological and geographic information.

The main factor for some of the outlying samples in the validation results was mainly high reflectance in ch01 due to the influence of thin clouds that were not screened by the cloud mask, which led to overestimations. In the analysis with separate test data, the models showed higher accuracy (R^2 value of 0.92 and RMSE of 0.085) when compared to the accuracy of the previous studies. The bias between predicted FMF and AERONET FMF showed that models tended to overestimate at very low values, which was found to mostly come from Beijing regions. This was likely attributed to the effect of high surface reflectance and cloud contamination in urban areas, which was consistent with the cross-validation results.

The most important feature for AOD estimation was ch01 reflectance, which means that AOD retrieval is most affected by the reflectance from aerosols. While the ch01 feature also contributed most to the FMF retrieval, other meteorological features were also considered important. It highlights the need to consider meteorological effects and geographic information in addition to spectral information from satellite channels for more accurate aerosol estimation. The role of the channel difference feature (i.e., TOA reflectance – minimum reflectance over the past 30 days), which represents the influence of surface reflectance, was important especially in bright desert regions. Models without the channel difference features tended to predict lower AOD, especially for regions with low aerosol loadings, as underlying surface reflectance is not taken into consideration. Therefore, it suggests that the channel difference features should be considered in such aerosol retrievals to account for the effect of surface reflectance.

In terms of the spatial distribution of AOD and FMF, a distinct aerosol plume was generally well detected by the models. However, Beijing regions, where moderate to high AOD values are often observed, were falsely screened by cloud masking, and thus aerosol information in such urban areas was not properly accounted for by the machine learning models. This suggests that more accurate cloud making would be needed for better aerosol retrievals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

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-SP2021-01-02-061), by the Korea Environment Industry & Technology Institute (KEITI) through Digital Infrastructure Building Project for Monitoring, Surveying and Evaluating the Environmental Health, funded by Korea Ministry of Environment (MOE) (2021003330001(NTIS: 1485017948)), and by the FRIEND (Fine Particle Research Initiative in East Asia Considering National Differences) Project, funded by the Ministry of Science and ICT (Grant No.: 2020M3G1A1114615). We thank the Atmospheric Radiation laboratory at the Yonsei University for providing GOCI aerosol products.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2021.11.016.

References

Bibi, H., Alam, K., Chishtie, F., Bibi, S., Shahid, I., Blaschke, T., 2015. Intercomparison of MODIS, MISR, OMI, and CALIPSO aerosol optical depth retrievals for four locations on the Indo-Gangetic plains and validation against AERONET data. Atmos. Environ. 111, 113–126.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., et al., 2013. Clouds and aerosols. Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, pp. 571–657.

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

Che, H., Xia, X., Zhu, J., Wang, H., Wang, Y., Sun, J., Zhang, X., Shi, G., 2015. Aerosol optical properties under the condition of heavy haze over an urban site of Beijing, China. Environ. Sci. Pollut. Res. 22 (2), 1043–1053.

Chen, X., de Leeuw, G., Arola, A., Liu, S., Liu, Y., Li, Z., Zhang, K., 2020. Joint retrieval of the aerosol fine mode fraction and optical depth using MODIS spectral reflectance over northern and eastern China: artificial neural network method. Remote Sens. Environ. 249, 112006. https://doi.org/10.1016/j.rse.2020.112006.

Cheng, L., Li, L., Chen, L., Hu, S., Yuan, L., Liu, Y., et al., 2019. Spatiotemporal variability and influencing factors of aerosol optical depth over the Pan Yangtze River Delta during the 2014–2017 period. Int. J. Environ. Res. Public Health 16, 3522.

Choi, M., Lim, H., Kim, J., Lee, S., Eck, T.F., Holben, B.N., Garay, M.J., Hyer, E.J., Saide, P.E., Liu, H., 2019. Validation, comparison, and integration of GOCI, AHI, MODIS, MISR, and VIIRS aerosol optical depth over East Asia during the 2016 KORUS-AQ campaign. Atmos. Meas. Tech. 12 (8), 4619–4641.

Choi, M., Kim, J., Lee, J., Kim, M., Park, Y.-J., Holben, B., et al., 2018. 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.

Choi, M., Kim, J., Lee, J., Kim, M., Park, Y.-J., Jeong, U., Kim, W., Hong, H., Holben, B., Eck, T.F., Song, C.H., Lim, J.-H., Song, C.-K., 2016. GOCI Yonsei Aerosol Retrieval (YAER) algorithm and validation during the DRAGON-NE Asia 2012 campaign. Atmos. Meas. Tech. 9 (3), 1377–1398.

Della Ceca, L.S., García Ferreyra, M.F., Lyapustin, A., Chudnovsky, A., Otero, L., Carreras, H., Barnaba, F., 2018. Satellite-based view of the aerosol spatial and temporal variability in the Córdoba region (Argentina) using over ten years of high-resolution data. ISPRS J. Photogramm. Remote Sens. 145, 250–267.
Dubovik, O., King, M.D., 2000. A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements. J. Geophys. Res.: Atmos. 105 (D16), 20673–20696.
Engström, A., Ekman, A.M., 2010. Impact of meteorological factors on the correlation between aerosol optical depth and cloud fraction. Geophys. Res. Lett. (37).
Fotiadi, A., Hatzianastassiou, N., Drakakis, E., Matsoukas, C., Pavlakis, K.G., Hatzidimitriou, D., Gerasopoulos, E., Mihalopoulos, N., Vardavas, I., 2006. Aerosol physical and optical properties in the Eastern Mediterranean Basin, Crete, from Aerosol Robotic Network data. Atmos. Chem. Phys. 6 (12), 5399–5413.
Giles, D.M., Sinyuk, A., Sorokin, M.G., Schafer, J.S., Smirnov, A., Slutsker, I., et al., 2019. Advancements in the Aerosol Robotic Network (AERONET) Version 3 database–automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol optical depth (AOD) measurements. Atmos. Meas. Tech. 12, 169–209.
Gumma, M.K., Thenkabail, P.S., Teluguntla, P.G., Oliphant, A., Xiong, J., Giri, C., Pyla, V., Dixit, S., Whitbread, A.M., 2020. Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud. GISci. Remote Sens. 57 (3), 302–322.
Gupta, P., Levy, R.C., Mattoo, S., Remer, L.A., Munchak, L.A., 2016. A surface reflectance scheme for retrieving aerosol optical depth over urban surfaces in MODIS Dark Target retrieval algorithm. Atmos. Meas. Tech. 9 (7), 3293–3308.
Holben, B.N., Eck, T.F., Slutsker, I., Tanré, D., Buis, J.P., Setzer, A., Vermote, E., Reagan, J.A., Kaufman, Y.J., Nakajima, T., Lavenu, F., Jankowiak, I., Smirnov, A., 1998. AERONET—a federated instrument network and data archive for aerosol characterization. Remote Sens. Environ. 66 (1), 1–16.
Hsu, N.C., Tsay, S.-C., King, M.D., Herman, J.R., 2004. Aerosol properties over bright-reflecting source regions. IEEE Trans. Geosci. Remote Sens. 42 (3), 557–569.
Huang, G., Chen, Y., Li, Z., Liu, Q., Wang, Y., He, Q., et al., 2020. Validation and accuracy analysis of the collection 6.1 MODIS aerosol optical depth over the westernmost city in China based on the sun-sky radiometer observations from SONET. Earth Space Sci. 7 e2019EA001041.
Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D., Liu, Y., 2018. Predicting monthly high-resolution PM2. 5 concentrations with random forest model in the North China Plain. Environ. Pollut. 242, 675–683.
Huttunen, J., Kokkola, H., Mielonen, T., Mononen, M.E.J., Lipponen, A., Reunanen, J., Lindfors, A.V., Mikkonen, S., Lehtinen, K.E.J., Kouremeti, N., Bais, A., Niska, H., Arola, A., 2016. Retrieval of aerosol optical depth from surface solar radiation measurements using machine learning algorithms, non-linear regression and a radiative transfer-based look-up table. Atmos. Chem. Phys. 16 (13), 8181–8191.
Jang, E., Im, J., Park, G.-H., Park, Y.-G., 2017. Estimation of fugacity of carbon dioxide in the East Sea using in situ measurements and Geostationary Ocean Color Imager satellite data. Remote Sens. 9 (8), 821. https://doi.org/10.3390/rs9080821.
Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., Kim, S., 2021. Estimation of surface-level NO2 and O3 concentrations using TROPOMI data and machine learning over East Asia. Environ. Pollut. 288, 117711. https://doi.org/10.1016/j.envpol.2021.117711.
Kaufman, Y.J., Tanré, D., Remer, L.A., Vermote, E.F., Chu, A., Holben, B.N., 1997. Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer. J. Geophys. Res.: Atmos. 102 (D14), 17051–17067.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inform. Process. Syst. 30, 3146–3154.
Khoshsima, M., Ahmadi-Givi, F., Bidokhti, A.A., Sabetghadam, S., 2014. Impact of meteorological parameters on relation between aerosol optical indices and air pollution in a sub-urban area. J. Aerosol Sci. 68, 46–57.
Kim, G., Lee, S., Im, J., Song, C.-K., Kim, J., Lee, M.-i., 2021. Aerosol data assimilation and forecast using Geostationary Ocean Color Imager aerosol optical depth and in-situ observations during the KORUS-AQ observing period. GISci. Remote Sens. 58 (7), 1175–1194.
Kittaka, C., Winker, D.M., Vaughan, M.A., Omar, A., Remer, L.A., 2011. Intercomparison of column aerosol optical depths from CALIPSO and MODIS-Aqua. Atmos. Meas. Tech. 4 (2), 131–141.
Kleidman, R.G., O'Neill, N.T., Remer, L.A., Kaufman, Y.J., Eck, T.F., Tanré, D., Dubovik, O., Holben, B.N., 2005. Comparison of Moderate Resolution Imaging Spectroradiometer (MODIS) and Aerosol Robotic Network (AERONET) remote-sensing retrievals of aerosol fine mode fraction over ocean. J. Geophys. Res.: Atmos. (110) 110 (D22). https://doi.org/10.1029/2005JD005760.
Lee, L.A., Carslaw, K.S., Pringle, K.J., Mann, G.W., Spracklen, D.V., 2011. Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. Atmos. Chem. Phys. 11 (23), 12253–12273.
Levy, R.C., Mattoo, S., Munchak, L.A., Remer, L.A., Sayer, A.M., Patadia, F., Hsu, N.C., 2013. The Collection 6 MODIS aerosol products over land and ocean. Atmos. Meas. Tech. 6 (11), 2989–3034.
Levy, R.C., Remer, L.A., Kleidman, R.G., Mattoo, S., Ichoku, C., Kahn, R., Eck, T.F., 2010. Global evaluation of the Collection 5 MODIS dark-target aerosol products over land. Atmos. Chem. Phys. 10 (21), 10399–10420.
Levy, R.C., Remer, L.A., Tanré, D., Mattoo, S., Kaufman, Y.J., 2009. Algorithm for remote sensing of tropospheric aerosol over dark targets from MODIS: Collections 005 and 051: Revision 2; Feb 2009. MODIS algorithm theoretical basis document.
Li, Z.Q., Xu, H., Li, K.T., Li, D.H., Xie, Y.S., Li, L., et al., 2018. Comprehensive study of optical, physical, chemical, and radiative properties of total columnar atmospheric aerosols over China: an overview of Sun-Sky Radiometer Observation Network (SONET) measurements. Bullet. Am. Meteorol. Soc. (99), 739–755.
Liang, T., Sun, L., Li, H., 2021. MODIS aerosol optical depth retrieval based on random forest approach. Remote Sens. Lett. 12 (2), 179–189. https://doi.org/10.1080/2150704X.2020.1842540.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.
Mai, B., Deng, X., Xia, X., Che, H., Guo, J., Liu, X., Zhu, J., Ling, C., 2018. Column-integrated aerosol optical properties of coarse-and fine-mode particles over the Pearl River Delta region in China. Sci. Total Environ. 622-623, 481–492. https://doi.org/10.1016/j.scitotenv.2017.11.348.
Martins, V.S., Novo, E.M.L.M., Lyapustin, A., Aragão, L.E.O.C., Freitas, S.R., Barbosa, C. C.F., 2018. Seasonal and interannual assessment of cloud cover and atmospheric constituents across the Amazon (2000–2015): insights for remote sensing and climate analysis. ISPRS J. Photogramm. Remote Sens. 145, 309–327.
Mu, Q., Liao, H., 2014. Simulation of the interannual variations of aerosols in China: role of variations in meteorological parameters. Atmos. Chem. Phys. 14 (18), 9597–9612.
Ng, D.H.L., Li, R., Raghavan, S.V., Liong, S.-Y., 2017. Investigating the relationship between aerosol optical depth and precipitation over Southeast Asia with relative humidity as an influencing factor. Sci. Rep. 7, 1–13.
Oros, D.R., Abas, M.R.B., Omar, N.Y.M.J., Rahman, N.A., Simoneit, B.R.T., 2006. Identification and emission factors of molecular tracers in organic aerosols from biomass burning: Part 3. Grasses. Appl. Geochem. 21 (6), 919–940.
Park, S., Lee, J., Im, J., Song, C.-K., Choi, M., Kim, J., Lee, S., Park, R., Kim, S.-M., Yoon, J., Lee, D.-W., Quackenbush, L.J., 2020. Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models. Sci. Total Environ. 713, 136516. https://doi.org/10.1016/j.scitotenv.2020.136516.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pham, T.D., Yokoya, N., Nguyen, T.T.T., Le, N.N., Ha, N.T., Xia, J., Takeuchi, W., Pham, T.D., 2021. Improvement of mangrove soil carbon stocks estimation in North Vietnam using sentinel-2 data and machine learning approach. GISci. Remote Sens. 58 (1), 68–87.
Pöschl, U., 2005. Atmospheric aerosols: composition, transformation, climate and health effects. Angew. Chem. Int. Ed. 44 (46), 7520–7540.
Rap, A., Scott, C.E., Reddington, C.L., Mercado, L., Ellis, R.J., Garraway, S., Evans, M.J., Beerling, D.J., MacKenzie, A.R., Hewitt, C.N., Spracklen, D.V., 2018. Enhanced global primary production by biogenic aerosol via diffuse radiation fertilization. Nat. Geosci. 11 (9), 640–644.
Reitz, O., Graf, A., Schmidt, M., Ketzler, G., Leuchner, M., 2021. Upscaling net ecosystem exchange over heterogeneous landscapes with machine learning. J. Geophys. Res. Biogeosci. 126 (2) e2020JG005814.
Remer, L.A., Kaufman, Y., Tanré, D., Mattoo, S., Chu, D., Martins, J.V., et al., 2005. The MODIS aerosol algorithm, products, and validation. J. Atmos. Sci. (62), 947–973.
Ryu, J.-H., Han, H.-J., Cho, S., Park, Y.-J., Ahn, Y.-H., 2012. Overview of geostationary ocean color imager (GOCI) and GOCI data processing system (GDPS). Ocean Sci. J. 47 (3), 223–233.
Sayer, A.M., Hsu, N.C., Bettenhausen, C., Jeong, M.-J., 2013. Validation and uncertainty estimates for MODIS Collection 6 "Deep Blue" aerosol data. J. Geophys. Res.: Atmos. 118 (14), 7864–7872.
She, L.u., Zhang, H.K., Li, Z., de Leeuw, G., Huang, B.o., 2020. Himawari-8 aerosol optical depth (AOD) retrieval using a deep neural network trained using AERONET observations. Remote Sens. 12 (24), 4125. https://doi.org/10.3390/rs12244125.
Shi, H., He, Q., Zhang, W., 2018. Spatial factor analysis for aerosol optical depth in metropolises in China with regard to spatial heterogeneity. Atmosphere 9 (4), 156. https://doi.org/10.3390/atmos9040156.
Shin, M., Kang, Y., Park, S., Im, J., Yoo, C., Quackenbush, L.J., 2020. Estimating ground-level particulate matter concentrations using satellite-based data: a review. GISci. Remote Sens. 57 (2), 174–189.
Tariq, S., Nawaz, H., Ul-Haq, Z., Mehmood, U., 2021. Investigating the relationship of aerosols with enhanced vegetation index and meteorological parameters over Pakistan. Atmos. Pollut. Res. 12 (6), 101080. https://doi.org/10.1016/j.apr.2021.101080.
Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Feichter, J., Fillmore, D., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Horowitz, L., Huang, P., Isaksen, I.S.A., Iversen, T., Kloster, S., Koch, D., Kirkevåg, A., Kristjansson, J.E., Krol, M., Lauer, A., Lamarque, J.F., Liu, X., Montanaro, V., Myhre, G., Penner, J.E., Pitari, G., Reddy, M. S., Seland, Ø., Stier, P., Takemura, T., Tie, X., 2007. The effect of harmonized emissions on aerosol properties in global models–an AeroCom experiment. Atmos. Chem. Phys. 7 (17), 4489–4501.
Unnithan, S.L.K., Gnanappazham, L., 2020. Spatiotemporal mixed effects modeling for the estimation of PM2. 5 from MODIS AOD over the Indian subcontinent. GISci. Remote Sens. 57 (2), 159–173.
Wang, Y., Yuan, Q., Li, T., Zhu, L., Zhang, L., 2021. Estimating daily full-coverage near surface O3, CO, and NO2 concentrations at a high spatial resolution over China based on S5P-TROPOMI and GEOS-FP. ISPRS J. Photogramm. Remote Sens. 175, 311–325.
Wei, J., Li, Z., Pinker, R.T., Wang, J., Sun, L., Xue, W., et al., 2021. Himawari-8-derived diurnal variations in ground-level PM 2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). Atmos. Chem. Phys. 21 (10), 7863–7880.

Yan, X., Li, Z., Shi, W., Luo, N., Wu, T., Zhao, W., 2017a. An improved algorithm for retrieving the fine-mode fraction of aerosol optical thickness, Part 1: Algorithm development. Remote Sens. Environ. 192, 87–97.

Yan, X., Shi, W., Li, Z., Li, Z., Luo, N., Zhao, W., Wang, H., Yu, X., 2017b. Satellite-based PM2. 5 estimation using fine-mode aerosol optical thickness over China. Atmos. Environ. 170, 290–302.

Yao, F., Wu, J., Li, W., Peng, J., 2019. A spatially structured adaptive two-stage model for retrieving ground-level PM2. 5 concentrations from VIIRS AOD in China. ISPRS J. Photogramm. Remote Sens. 151, 263–276.

Yeom, Jong-Min, et al., 2021. Estimation of the hourly aerosol optical depth from GOCI geostationary satellite data: deep neural network, machine learning, and physical models. IEEE Trans. Geosci. Remote Sens.

Yoon, S.-C., Kim, J., 2006. Influences of relative humidity on aerosol optical properties and aerosol radiative forcing during ACE-Asia. Atmos. Environ. 40 (23), 4328–4338.

Zhang, K., de Leeuw, G., Yang, Z., Chen, X., Su, X., Jiao, J., 2019. Estimating spatio-temporal variations of PM2. 5 concentrations using VIIRS-derived AOD in the Guanzhong Basin, China. Remote Sens. 11 (22), 2679. https://doi.org/10.3390/rs11222679.