Environment International 146 (2021) 106290

Contents lists available at ScienceDirect

Environment International

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

The ChinaHighPM_{10} dataset: generation, validation, and spatiotemporal variations from 2015 to 2019 across China

Jing Wei [a,b,*], Zhanqing Li [b,*], Wenhao Xue [a], Lin Sun [c], Tianyi Fan [a], Lei Liu [d], Tianning Su [b], Maureen Cribb [b]

[a] State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
[b] Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[c] College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao, China
[d] College of Earth and Environmental Sciences, Lanzhou University, Lanzhou, China

ARTICLE INFO

Communicated by: Hefa Cheng

Keywords:
Remote sensing
PM_{10}
AOD
1-km resolution
ChinaHighPM_{10}

ABSTRACT

Respirable particles with aerodynamic diameters ≤ 10 μm (PM_{10}) have important impacts on the atmospheric environment and human health. Available PM_{10} datasets have coarse spatial resolutions, limiting their applications, especially at the city level. A tree-based ensemble learning model, which accounts for spatiotemporal information (i.e., space-time extremely randomized trees, denoted as the STET model), is designed to estimate near-surface PM_{10} concentrations. The 1-km resolution Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol product and auxiliary factors, including meteorology, land-use cover, surface elevation, population distribution, and pollutant emissions, are used in the STET model to generate the high-resolution (1 km) and high-quality PM_{10} dataset for China (i.e., ChinaHighPM_{10}) from 2015 to 2019. The product has an out-of-sample (out-of-station) cross-validation coefficient of determination (CV-R^2) of 0.86 (0.82) and a root-mean-square error (RMSE) of 24.28 (27.07) μg/m^3, outperforming most widely used models from previous related studies. High levels of PM_{10} concentration occurred in northwest China (e.g., the Tarim Basin) and the Northern China Plain. Overall, PM_{10} concentrations had a significant declining trend of 5.81 μg/m^3 per year ($p < 0.001$) over the past five years in China, especially in three key urban agglomerations. The ChinaHighPM_{10} dataset is potentially useful for future small- and medium-scale air pollution studies by virtue of its higher spatial resolution and overall accuracy.

1. Introduction

In recent years, several acute air pollution episodes have occurred in mainland China due partially to urban expansion and industrial development (Chan and Yao, 2008; Ji et al., 2012; Xu et al., 2013; Sun et al., 2016; Guo et al., 2017; Su et al., 2018), mainly involving coarse and fine particulate matter with aerodynamic diameters of no more than 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$). While PM_{10} mainly comes from nature, e.g., dust, soil, and sea salt, anthropogenic activities also play an important role in the emission of PM_{10}, such as construction-generated dust, insufficient combustion of fossil fuels, and discharge of industrial residues (Bi et al., 2007; Rohde and Muller, 2015; Wei et al., 2019a, b). PM_{10} is thus of great concern to both the atmospheric environment (Choi et al., 2008; Qu et al., 2010) and human health, given its potential contribution to cancer, respiratory diseases, and heart diseases, especially in developing countries like China (Bartell et al., 2013; Xu et al., 2013; Liu et al., 2019a).

The Chinese Ministry of Environmental Protection has thus established a national ground-based monitoring network to monitor typical air pollutants (e.g., PM_{10}, $PM_{2.5}$, ozone, and sulfur dioxide) in real time (Guo et al., 2009). However, these monitoring stations are sparsely and non-uniformly distributed across mainland China with large gaps in coverage. Satellite remote sensing has the advantage of providing complete and uniform coverage. Official aerosol optical depth (AOD) products have been derived from multiple satellite sensors, including, among many others, the Multi-angle Imaging SpectroRadiometer (Garay et al., 2020), the Moderate-resolution Imaging Spectroradiometer (MODIS; Levy et al., 2013; Wei et al., 2019c), the Visible Infrared

* Corresponding authors at: State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China (J. Wei).
E-mail addresses: weijing_rs@163.com (J. Wei), zli@atmos.umd.edu (Z. Li).

https://doi.org/10.1016/j.envint.2020.106290
Received 14 May 2020; Received in revised form 8 October 2020; Accepted 16 November 2020
Available online 11 December 2020

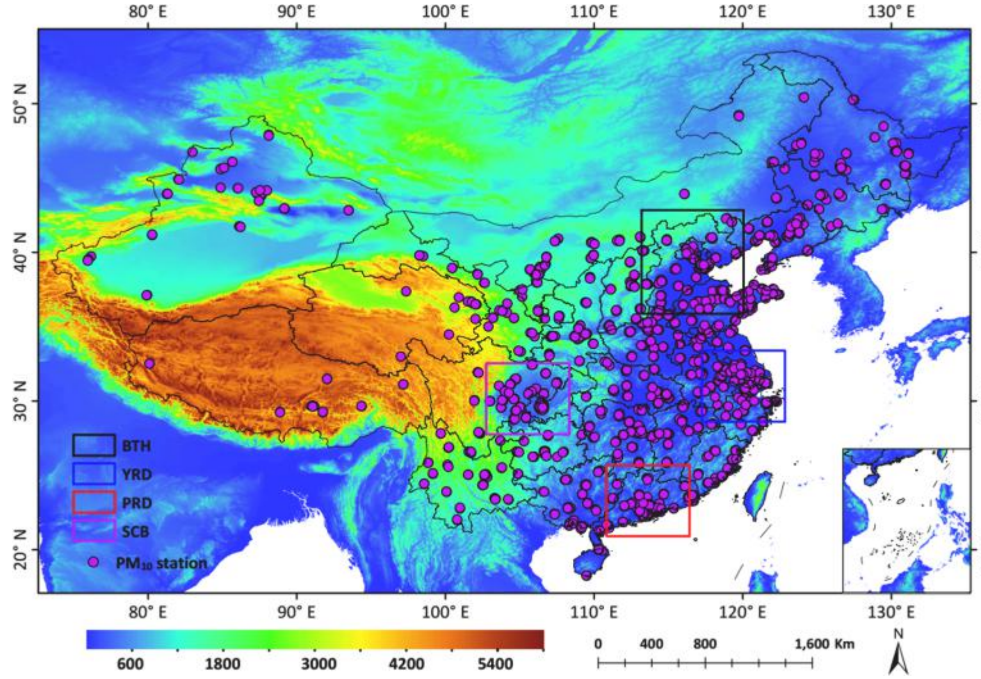

Fig. 1. Spatial distribution of surface PM_{10} monitoring stations in China (purple dots). The background map shows digital elevation model data (unit: m). Colored boxes outline four regions of interest: the Beijing-Tianjin-Hebei (BTH) region, the Yangtze River Delta (YRD) region, the Pearl River Delta (PRD) region, and the Sichuan Basin (SCB). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Imaging Radiometer (Hsu et al., 2019; Jackson et al., 2013), and the Himawari-8 Advanced Himawari Imager (Yoshida et al., 2018; Su et al., 2020; Zhang et al., 2019). A positive relationship between AOD and near-surface PM concentrations has been shown (Guo et al., 2009) despite uncertainties due to various sources of the retrievals (Li et al., 2009). As a result, these products have been employed to estimate near-surface PM concentrations (including PM_1, $PM_{2.5}$, and PM_{10}) from regional to global scales (Ma et al., 2014; Franklin et al., 2017; Su et al., 2017; Chen et al., 2018; Zang et al., 2018; Wei et al., 2019a, b; Yao et al., 2019). At present, satellite-based PM estimation methods fall into four main categories: physical models (Koelemeijer et al., 2006; Emili et al., 2010; Wang et al., 2014; Zhang and Li, 2015), chemical models (Ghotbi et al., 2016), statistical regression models (Nordio et al., 2013; Ma et al., 2014; You et al., 2015), and artificial intelligence (Zaman et al., 2017; Chen et al., 2018; Zang et al., 2018; Zhang et al., 2018; Wei et al., 2019a, b).

For near-surface PM_{10} species, an increasing number of studies have been carried out around the world over the years. Koelemeijer et al. (2006) and Emili et al. (2010) used the traditional physical method to derive 10-km-resolution PM_{10} concentrations over Europe from satellite AODs by correcting the planetary boundary layer height (BLH) and relative humidity (RH). Benas et al. (2013) used a nonlinear ACE algorithm to predict PM_{10} concentrations in Athens, Greece, at a 10-km resolution. Nordio et al. (2013) applied the linear mixed-effect (LME) model to estimate PM_{10} concentrations in Lombardy, northern Italy, at a 10-km resolution. Sotoudeheian and Arhami (2014) developed linear and non-linear multi-regression models to derive PM_{10} concentrations in Tehran, Iran, at 10-km and 17.6-km resolutions. Beloconi et al. (2016) estimated PM_{10} and $PM_{2.5}$ concentrations in London, UK, using the mixed-effect model with day-specific random effects at a 1-km resolution. Ghotbi et al. (2016) used the Weather Research and Forecasting model to estimate PM_{10} concentrations at a 3-km resolution in Tehran, Iran. Zaman et al. (2017) adopted the multiple linear regression (MLR) and artificial neural network models to derive 10-km-resolution PM_{10} concentrations in Malaysia. Stafoggia et al. (2019) developed a spatiotemporal land-use random-forest model to estimate daily PM_{10} and $PM_{2.5}$ concentrations in Italy at a 1-km resolution.

Many satellite-based PM_{10} estimation studies with a focus on China have also been performed due to the more serious air pollution problem in that country. Wang et al. (2014) developed an empirical method by correcting the RH to estimate 10-km-resolution PM_{10} concentrations in Beijing, China. You et al. (2015) applied the geographically weighted regression (GWR) model to estimate PM_{10} concentrations at a 10-km resolution in northwestern China, then developed a nonlinear empirical model to derive PM_{10} concentrations (10 km) in a semi-arid area, i.e., Xi'an City in China (You et al., 2016). Meng et al. (2015) used the LME model to generate 3-km-resolution PM_{10} maps for Shanghai, China. Zhang et al. (2016) combined a physical model and the GWR model to derive 3-km-resolution PM_{10} concentrations across China. Zhang et al. (2018) developed a spatiotemporal land-use regression (LUR) model to estimate monthly PM_{10} concentrations at a 10-km resolution from 2014 to 2016 in China. Chen et al. (2018) employed the random forest (RF) model to predict historical PM_{10} records at a horizontal resolution of $0.1^\circ \times 0.1^\circ$ from 2005 to 2016 in China.

Despite these efforts, further improvements can be made in estimating near-surface PM_{10} concentrations in China. First, traditional physical, chemical, and statistical regression models generally have less data mining abilities than do machine learning methods. The capability of the latter can be further improved by accounting for the spatiotemporal continuity (Li et al., 2017b; Wei et al., 2019a). Second, most previous studies employed MODIS AOD products generated from the Dark Target (DT) or Deep Blue (DB) aerosol algorithms at spatial resolutions of 3–10 km. The coarse-resolution PM_{10} dataset generated thus has limited applications in medium- or small-scale areas such as urban regions where air pollution is much more locally concentrated than in rural regions. Last, PM_{10} is less intensively studied than $PM_{2.5}$ in China, possibly because coarser particles have less impact on human health for the same mass concentration. However, the influence of PM_{10} on the atmospheric environment cannot be ignored. Therefore, a high-resolution, high-quality PM_{10} dataset is of potential use.

To overcome these limitations, we have adopted here a more accurate ensemble learning approach by integrating spatiotemporal information, resulting in a method called the space-time extremely randomized trees (STET) (Wei et al., 2020). This method was used to establish robust PM_{10}-AOD relationships. A high-quality PM_{10} dataset at a spatial resolution of 1 km covering China was finally generated from the STET model, i.e., ChinaHighPM_{10}, using the newly released MODIS 1-km Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol product (Lyapustin et al., 2018), along with meteorological, land use, topography, and population data as input. Section 2 introduces the data sources and integration, model development, and evaluation approaches. Section 3 evaluates the satellite-derived PM_{10} estimates from 2015 to 2019 at different spatiotemporal scales, then investigates their spatiotemporal variations across China. Comparisons between the model performance of the STET model and that of traditional models presented in previous similar studies are also discussed. Section 4 summarizes the study.

2. Data and method

2.1. Data sources and integration

2.1.1. PM_{10} in situ data

PM_{10} near the ground has been monitored across China by China's National Environmental Monitoring Center (CNEMC) at 1480 stations in 2015 then increased to 1605 stations in 2019. They are more evenly and densely distributed in eastern China than in western China (Fig. 1). Here, hourly PM_{10} measurements are first checked to remove outliers caused by instrument malfunction (Guo et al., 2009). Then valid hourly measurements are averaged to obtain daily means in a year for each monitoring station in China.

2.1.2. MAIAC AOD product

The Collection 6 MAIAC aerosol product at a 1-km spatial resolution (Lyapustin et al., 2018) is selected as a key input to estimate near-surface PM_{10} concentrations during 2015–2019 covering the whole of China. AOD data at 550 nm were derived from the MODIS onboard the Terra (10:30 am) and Aqua (1:30 pm) platforms passing the recommended data quality flags (i.e., $QA_{CloudMask}$ = Clear and $QA_{AdjacencyMask}$ = Clear). They are averaged to obtain daily means, improving the spatial coverage by 23–27%. Note that the diurnal mean may, at times, be derived from a single sample when the other is unavailable due to clouds or for other reasons. In addition, this can also increase the number of PM_{10}-AOD matchups (Wei et al., 2019a).

2.1.3. ERA5 reanalysis product

Most previous studies have mainly used the National Centers for Environmental Prediction (NCEP), Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2), and European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis Interim (ERA-Interim) atmospheric reanalysis products to provide meteorological observations. However, they are generated at much coarser spatial resolutions (e.g., 0.5°×0.5° and 2.5°×2.5°) and low temporal resolutions (e.g., 3–6 h) (Sun et al., 2018). By contrast, in our study, the latest release of the ERA5 atmospheric reanalysis product with high spatial (i.e., 0.1°×0.1°) and temporal (i.e., 1 h) resolutions are employed (Copernicus Climate Change Service, 2017). It includes the ERA5-Land hourly dataset beginning in January 1981 to the present at a horizontal resolution of 0.1°×0.1° (released on 12 July 2019) and ERA5 global hourly data on single or multiple pressure levels beginning in 1979 to the present at a horizontal resolution of 0.25°×0.25° (released on 14 June 2018). Here, five ERA5-Land variables, i.e., evapotranspiration (ET), temperature at a height of 2 m (TEM), surface pressure (SP), and the u- and v-components of wind at a height of 10 m (WU and WV, respectively), and two ERA5 global variables, i.e., BLH and RH, are selected. Meteorological observations are averaged from 10:00 am to 2:00 pm to obtain daily mean values.

2.1.4. Auxiliary data

Auxiliary data, including land use, topography, human population distribution, and pollution discharge, which have potential effects on PM_{10} pollution, are considered. Two land-use-related indices, i.e., the MODIS annual land-use cover and monthly Normalized Difference Vegetation Index (NDVI) products, and the Shuttle Radar Topography Mission (SRTM) surface elevation (DEM) data are also employed. The LandScan^{TM} annual population distribution (POP; Dobson et al., 2000) product and the monthly pollutant emissions contributing to coarse particles (PM) from agriculture, industry, power, residential, and transportation from the multi-resolution emission inventory for China (MEIC; Zhang et al., 2007; Li et al., 2017a) are additional inputs. The emissions generally have much smaller-scale variations than meteorological variables. Similar to previous studies, all coarser-resolution meteorological and auxiliary data are resampled to the same 1-km spatial resolution using the bilinear interpolation method to be consistent with the AOD product. Table 1 summarizes the data sources used in our study.

2.2. Model introduction and validation

2.2.1. Space-time extremely randomized trees

A typical tree-based ensemble learning method, called the extremely randomized trees (extra-trees, ERT), was employed here (Geurts et al., 2006). The ERT model consists of hundreds to thousands of decision trees that can be used for addressing regression and classification issues. It further strengthens the randomization of attribute selection and node splitting and can effectively reduce the model variance, differing from other popular models like the decision tree (DCT) and RF (Breiman, 2001).

More importantly, compared with deep learning and other traditional machine learning approaches, ensemble learning methods have unique advantages. The essential one is that they are not sensitive to multivariate collinearity variables and can process a large volume of

Table 1
Summary of data sources used in this study.

Dataset	Variable	Content	Unit	Spatial Resolution	Temporal Resolution	Data Source
PM_{10}	PM_{10}	PM_{10}	μg/m^3	In situ	Hourly	CNEMC
AOD	AOD	MAIAC AOD	–	1 km × 1 km	Daily	MCD19A2
Meteorology	ET	Evapotranspiration	mm	0.1°×0.1°	Hourly	ERA5
	TEM	2-m temperature	K	0.1°×0.1°	Hourly	
	SP	Surface pressure	hPa	0.1°×0.1°	Hourly	
	WU	10-m u-component of wind	m/s	0.1°×0.1°	Hourly	
	WV	10-m v-component of wind	m/s	0.1°×0.1°	Hourly	
	BLH	Boundary layer height	m	0.25°×0.25°	Hourly	
	RH	Relative humidity	%	0.25°×0.25°	Hourly	
Land cover	LUC	Land-use cover	–	500 m × 500 m	Yearly	MCD12Q1
	NDVI	NDVI	–	1 km × 1 km	Monthly	MOD13A3
Topography	DEM	Surface elevation	m	90 m × 90 m	–	SRTM
Population	POP	Ambient population	–	1 km × 1 km	Yearly	LandScan^{TM}
	PM	PM emission	Mg/grid	0.25°×0.25°	Monthly	MEIC

input data without the need for a reduction in data dimensionality. That is to say, they can mine valuable information and discard useless information internally during the model building, unlike traditional models (Breiman, 2001; Geurts et al., 2006). Therefore, all the above-mentioned variables, including MAIAC AOD, meteorology, pollution distribution and emission, land cover, and topography, are selected, together with PM_{10} ground measurements, as inputs into the ERT model.

Furthermore, considering that PM_{10} concentrations exhibit noticeable spatiotemporal heterogeneities, a new space-time tree-based ensemble learning approach was developed, i.e., the STET model (Wei et al., 2020), by incorporating spatial and temporal information into the original ERT model to improve the overall estimation accuracy of PM_{10} estimates. The space term includes the linear latitude (X) and longitude (Y) of one point, and five additional non-linear spatial distance fields, i.e., distances to the upper-left (D_1), upper-right (D_2), lower-left (D_3), lower-right (D_4) corners, and the center (D_5) of a circumscribed rectangle in the study area (Wei et al., 2021). They can be jointly used to explicitly describe the autocorrelated spatial position of one point in space (Krumbein, 1959; Behrens et al., 2018). These distances are great-circle distances between two points on a sphere, calculated using the Haversine approach (Eq. (1)). In addition, unlike other traditional statistical regression or artificial intelligence methods, the tree-based machine learning methods we used are supervised classification methods, whose basic unit is the decision tree. Although the distances can be calculated by longitude and latitude, all the input variables are totally independent, and no other combination operation is performed in the node splitting during the model training (Breiman, 2001; Geurts et al., 2006). The simplified time term is represented by the day of the year (DOY), which is used to mark each row of data records of one point in space on different days in a year during the tree-based ensemble model training because PM_{10} concentrations vary with location and day of the year (Wei et al., 2021). The PM_{10}-AOD relationship (Eq. (2)) can then be established using the STET model:

$$DIS_{i,j,t} = 2*r*\mathrm{asin}\left\{sqrt\left[sin^2\left(\frac{Lat_{i,j}-Lat_0}{2}\right) + cos(Lat_{i,j})cos(Lat_0)sin^2\left(\frac{Lon_{i,j}-Lon_0}{2}\right)\right]\right\} \tag{1}$$

$$PM_{10(i,j,t)} = f_{STET}[AOD_{i,j,t}, BLH_{i,j,t}, DEM_{i,j,t}, EP_{i,j,t}, LUC_{i,j,t}, NDVI_{i,j,t}, PM_{i,j,t}, POP_{i,j,t}, RH_{i,j,t}, SP_{i,j,t}, TEM_{i,j,t}, WU_{i,j,t}, WV_{i,j,t}, Lon_{i,j,t}, Lat_{i,j,t}, DIS_{i,j,t}, DOY_{i,j}] \tag{2}$$

where $Lat_{i,j}$, $Lon_{i,j}$ and Lat_0, Lon_0 denote the latitudes and the longitudes of one point $P(i, j)$ and the corner or centre $P(0, 0)$ of a rectangle in space, r represents the earth's radius (r = 6371 km), and DOY and N refer to the ith day and the total number of days of the year, respectively.

For the STET model, daily PM_{10} ground-based measurements, AOD, along with 12 auxiliary factors, are first spatiotemporally collocated at each PM_{10} monitoring station in each year and are used to form the data samples. Second, a training set (n) is randomly selected from all data samples (N), and m features are randomly selected from all features (M) without replacement. Then a split (s^*) is selected among all generated splits (S) according to the calculated scores (s^*, S), and an extremely randomized tree is built based on the Classification And Regression Tree (CART) algorithm. Last, the above steps are repeated to construct numerous extremely randomized trees as weak classifiers, which are then combined to form a strong classifier. Geurts et al. (2006) provide

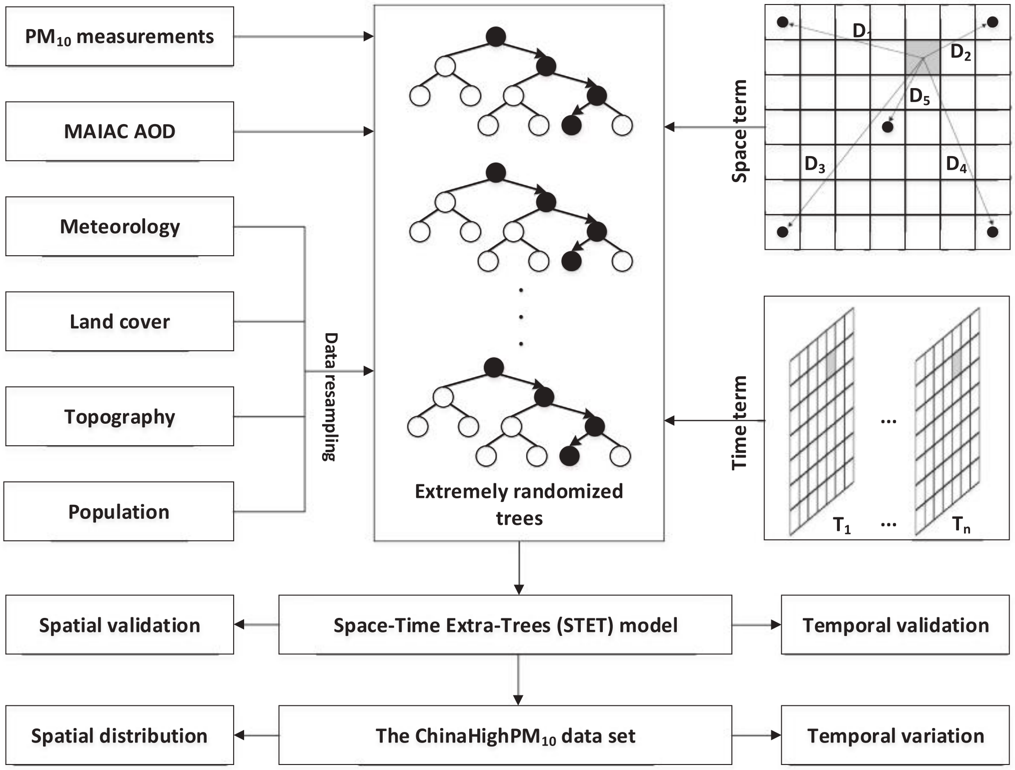

Fig. 2. Flowchart describing how the ChinaHighPM_{10} dataset is generated.

detailed information about the ERT model. Fig. 2 shows a flowchart describing how the ChinaHighPM_{10} dataset was generated in our study.

2.2.2. Evaluation and analysis approach

In our study, two typical independent 10-fold cross-validation (CV) approaches (Rodriguez et al., 2010) based on all data samples (i.e., out-of-sample validation; Ma et al., 2014) and PM_{10} monitoring stations (i.e., out-of-station validation; Wei et al. (2020) are selected to validate the overall accuracy and spatial prediction ability of models, respectively. The data samples or monitoring stations are divided into ten random subsets, where nine (one) of the subsets are used for training (validation), in turn, ten times (Rodriguez et al., 2010). In addition, traditional statistical metrics, including the regression line, the coefficient of determination (R^2), the mean absolute error (MAE), and the root-mean-square error (RMSE), are used to evaluate the overall accuracy and uncertainty. Daily PM_{10} maps are used to generate monthly, seasonal, and annual maps using the spatiotemporally average method. Temporal trends are calculated using the linear regression method based on deseasonalized monthly PM_{10} anomalies, and the trend significance is validated using the two-side test approach (Wei et al., 2019d).

3. Results and discussion

3.1. Validation against ground measurements

3.1.1. Spatial-scale validation

Fig. 3 illustrates the out-of-sample CV results of daily PM_{10} estimates in China from 2015 to 2019 using the STET model. The PM_{10}-AOD matchups are densely distributed on both sides of the 1:1 line in each year, especially in the PM_{10} concentration range of 0 to 200 μg/m^3, which has the largest data density. Despite some differences, satellite-derived PM_{10} concentrations agree well with the corresponding observations (CV-R^2 = 0.83–0.87), with strong slopes of 0.78–0.82 and small intercepts of 15.2–22.7 μg/m^3 among different years. There are overall low uncertainties with small RMSEs of 19.7–28.4 μg/m^3 and MAEs of 11.5–17.9 μg/m^3. Overall, our model shows high accuracy in deriving daily PM_{10} concentrations, with a high out-of-sample CV-R^2 of 0.86 during 2015–2019 across China, and the average RMSE and MAE values are 24.28 μg/m^3 and 14.52 μg/m^3, respectively.

Out-of-station CV results of daily PM_{10} estimates in China from 2015 to 2019 using the STET model offer similar conclusions (Fig. 4). The STET-model-predicted daily PM_{10} concentrations correlate well with surface observations (CV-R^2 = 0.79–0.84; Slope = 0.75–0.80) with small estimation uncertainties (RMSE = 21.8–31.4 μg/m^3, MAE = 13.4–20.6 μg/m^3) among different years. Overall, the out-of-station CV-R^2 is 0.82, and the RMSE and MAE are 27.07 μg/m^3 and 16.86 μg/m^3 during 2015–2019 in China, respectively. The station-based performance did not deteriorate much, with small differences in most evaluation metrics compared to the sample-based performance in every single year in China. These results suggest that the STET model has a high spatial prediction ability and can estimate PM_{10} concentrations well for areas in China without monitoring stations.

Table 2 provides sample- and station-based CV results of daily PM_{10} estimates derived from the STET model during 2015–2019 at the regional scale in China. Eastern China is a region where the STET model is overall more accurate with a stronger prediction ability than in western China, with a higher sample-based CV-R^2 of 0.88 and station-based CV-R^2 of 0.85 (0.82 and 0.76, respectively, in western China). The main reason is that the number of PM_{10} ground monitoring stations in eastern China is about four times that in western China. However, the STET model performs differently in different smaller regions. For example, it can well estimate and predict daily PM_{10} concentrations in

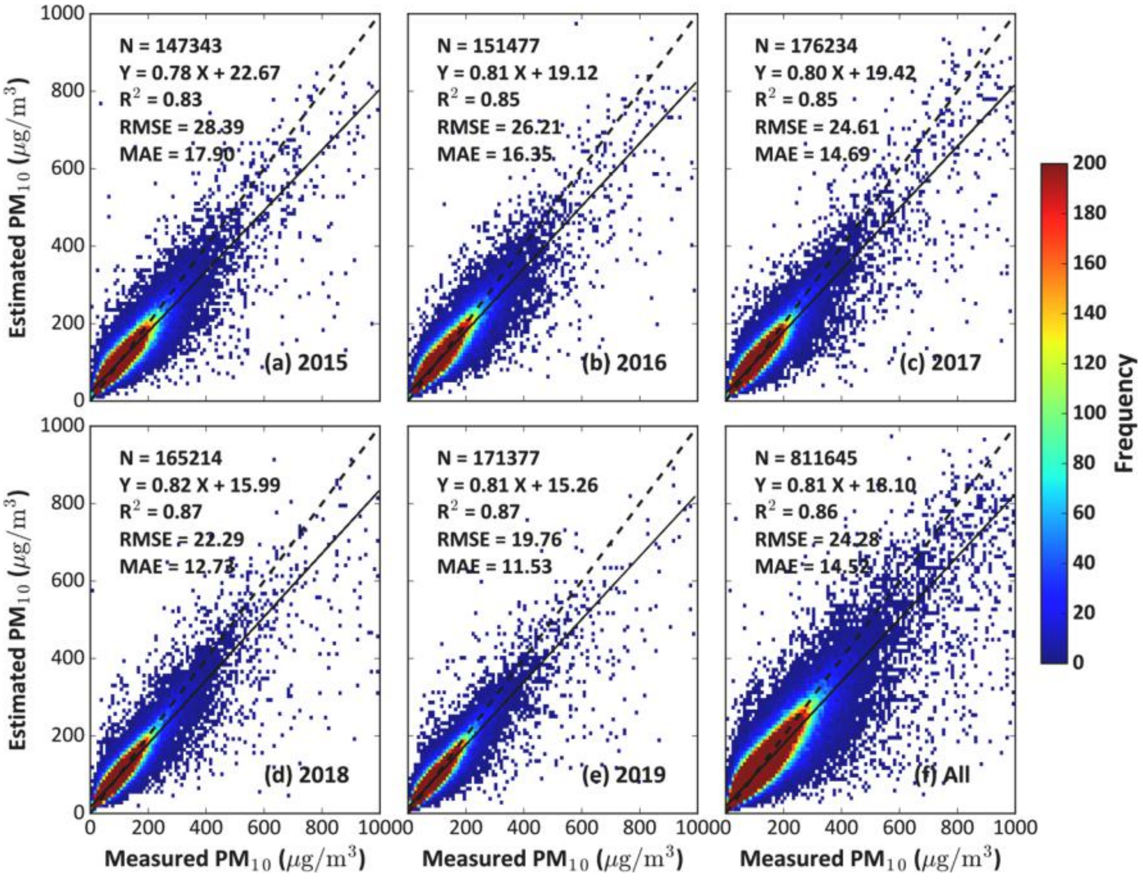

Fig. 3. Density scatter plots of out-of-sample cross-validation results in daily PM_{10} estimates from 2015 to 2019 across China.

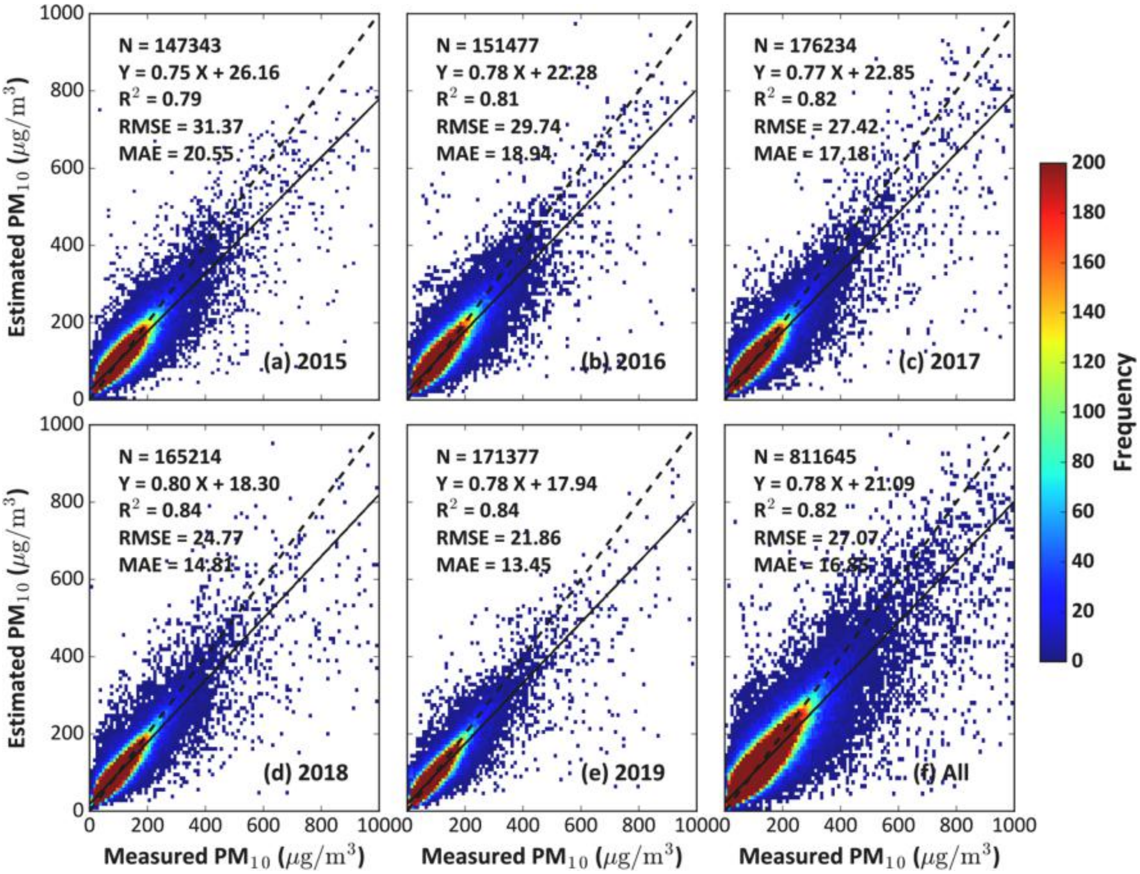

Fig. 4. Same as Fig. 3 but for out-of-station cross-validation results.

the Beijing-Tianjin-Hebei (BTH) region (i.e., sample-based CV-$R^2 = 0.89$ and station-based CV-$R^2 = 0.87$). Here, more severe air pollution with high PM_{10} levels leads to relatively large RMSE and MAE values. By contrast, there are overall low correlations (i.e., sample- and station-based CV-$R^2 = 0.84$ and 0.79, respectively) between PM_{10} estimates and measurements in the Pearl River Delta (PRD) region, with smaller estimation uncertainties due to lower PM_{10} levels. In general, the STET model shows similar performance with close statistical metrics between estimated and measured PM_{10} concentrations in the Yangtze River Delta (YRD) and the Sichuan Basin (SCB) regions.

Fig. 5 illustrates the sample- and station-based accuracies and uncertainties of the STET model in daily PM_{10} estimates from 2015 to 2019 at each monitoring station in China. Concerning the out-of-sample validation, the STET model can well estimate daily PM_{10} concentrations at most stations, where more than 86%, 90%, and 91% of the stations have CV-R^2 values > 0.7, RMSEs < 30 μg/m^3, and MAEs < 20 μg/m^3, respectively. Poor accuracies with large uncertainties are mainly found at several monitoring stations located in western China. The station-based validation results show similar spatial patterns as the sample-based validation results at the site scale across China. In general, the PM_{10} predictions are highly correlated to ground measurements with small CV-R^2, RMSE, and MAE values at 81%, 85%, and 81% of the stations in China, respectively. The STET model shows a stronger spatial ability in predicting daily PM_{10} concentrations at most stations located in eastern China than in western China. The large difference in the number of monitoring stations between western and eastern China, as well as both natural and human influences, may explain this.

3.1.2. *Temporal-scale performance*

Fig. 6 shows the sample- and station-based CV results of all estimated daily PM_{10} concentrations from all monitoring stations on each DOY from 2015 to 2019 across China. There are abundant data samples ranging from 605 to 4046, with an average of 2218 samples for each day from 2015 to 2019. The STET model can well capture PM_{10} concentrations on most days, with 95% and 75% of the days showing high sample- and station-based CV-R^2 values, respectively, greater than 0.7. However, the STET model performs less well in the middle of the year, with overall low sample- and station-based CV-R^2 values. The main reason is that there is a large data gap in the aerosol product caused by the presence of summertime clouds, seriously limiting the training ability of the model. In addition, the PM_{10} estimates (predictions) show low estimation uncertainties on more than 87% (79%) and 95% (86%) of the days, with small RSMEs and MAEs < 30 μg/m^3 and 20 μg/m^3, respectively. The estimation uncertainties are generally large at the beginning and the end of the year, mainly due to frequent sandstorms and a large number of pollution emissions in spring and winter in northern China.

Satellite-based monthly synthetic PM_{10} estimates were also validated

Table 2
Statistics of out-of-sample and out-of-station cross-validation results in daily PM_{10} estimates in China and in each region of interest from 2015 to 2019. Units for RMSE and MAE values are μg/m^3.

Region	N	Out-of-sample validation			Out-of-station validation		
		R^2	RMSE	MAE	R^2	RMSE	MAE
ECHN	609,363	0.88	20.32	13.15	0.85	22.13	14.81
WCHN	202,282	0.82	33.50	18.64	0.76	38.27	23.00
BTH	122,927	0.89	24.81	15.99	0.87	26.28	17.28
YRD	72,097	0.86	16.40	11.32	0.83	17.64	12.36
PRD	31,910	0.84	12.06	8.44	0.79	13.69	9.69
SCB	27,588	0.86	17.89	12.68	0.83	19.81	14.24

BTH: Beijing-Tianjin-Hebei; ECHN: eastern China; PRD: Pearl River Delta; SCB: Sichuan Basin; WCHN: western China; YRD: Yangtze River Delta.

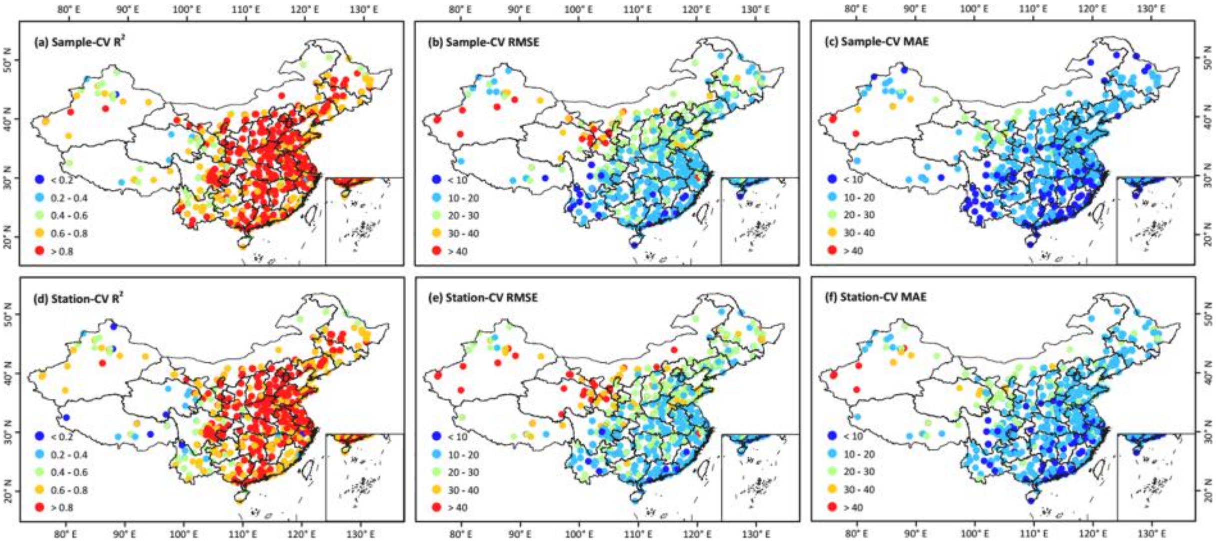

Fig. 5. Spatial distributions of (a-c) out-of-sample and (d-f) out-of-station cross-validation (CV) results in daily PM_{10} estimates at each monitoring station from 2015 to 2019 across China. Units for RMSE and MAE values are μg/m^3.

and compared with the corresponding monthly PM_{10} ground measurements for each year in China (Fig. 7). Over the years considered, monthly PM_{10} estimates have high accuracies, with coefficients of determination ranging from 0.92 to 0.95 and overall small estimation uncertainties (i.e., RMSE = 8.4–13.7 μg/m^3, MAE = 6.1–9.8 μg/m^3). A total of 59,079 monthly PM_{10} matchups were collected from 2015 to 2019 across China (Fig. 7f) which were highly consistent (R^2 = 0.94), and with average RMSE and MAE values of 11.07 μg/m^3 and 7.87 μg/m^3, respectively. These results suggest that the monthly synthetic data also yield high accuracies and can well capture the spatiotemporal variations of PM_{10} pollution across China.

3.2. Spatiotemporal characteristics

Here, the STET model is applied to generate a daily high-resolution (1 km) and high-quality PM_{10} dataset for China (i.e., ChinaHighPM_{10}) from 2015 to 2019. Daily PM_{10} maps are then synthesized to obtain the monthly, seasonal, and annual mean PM_{10} maps, which are used to explore PM_{10} spatiotemporal characteristics across China.

3.2.1. Spatial coverage and distribution

Fig. 8 shows satellite-derived 1-km-resolution (≈ 0.01° × 0.01°) annual mean PM_{10} maps from 2015 to 2019 across China, and Table 3 summarizes the statistics for China and each region of interest. The STET model can generate spatially continuous PM_{10} maps and can cover most areas of China with an average spatial coverage of 99%. Although there are some differences in the spatial patterns among the years, the overall pollution level appears to have gradually decreased over the years. Note that there are large differences in PM_{10} concentrations over the years in the Tibetan Plateau, mainly due to the small number of sparsely

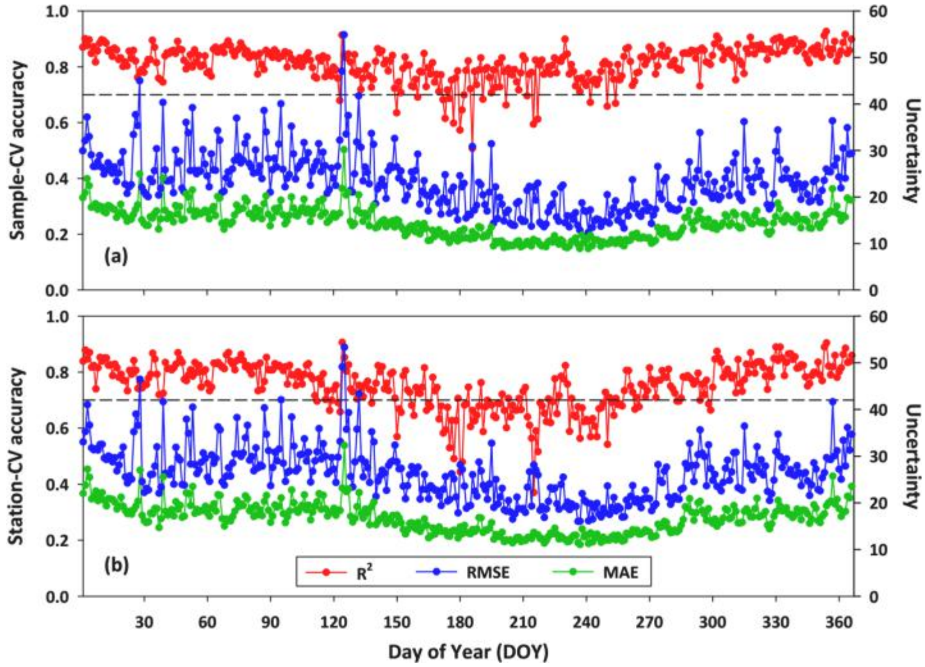

Fig. 6. Time series of the daily performance in PM_{10} (a) estimates and (b) predictions using the STET model from 2015 to 2019 across China. Units for RMSE and MAE values are μg/m^3.

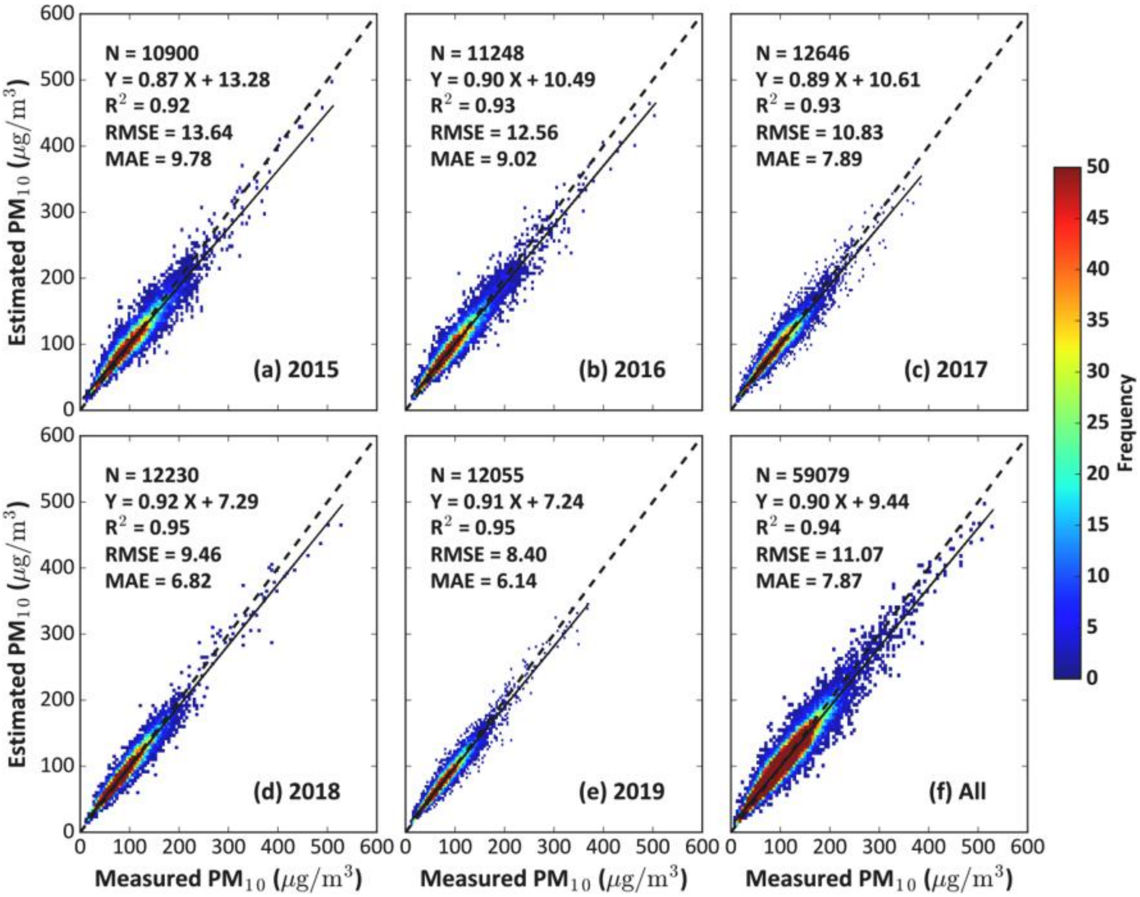

Fig. 7. Validation of monthly synthetic PM_{10} estimates against ground measurements from 2015 to 2019 across China.

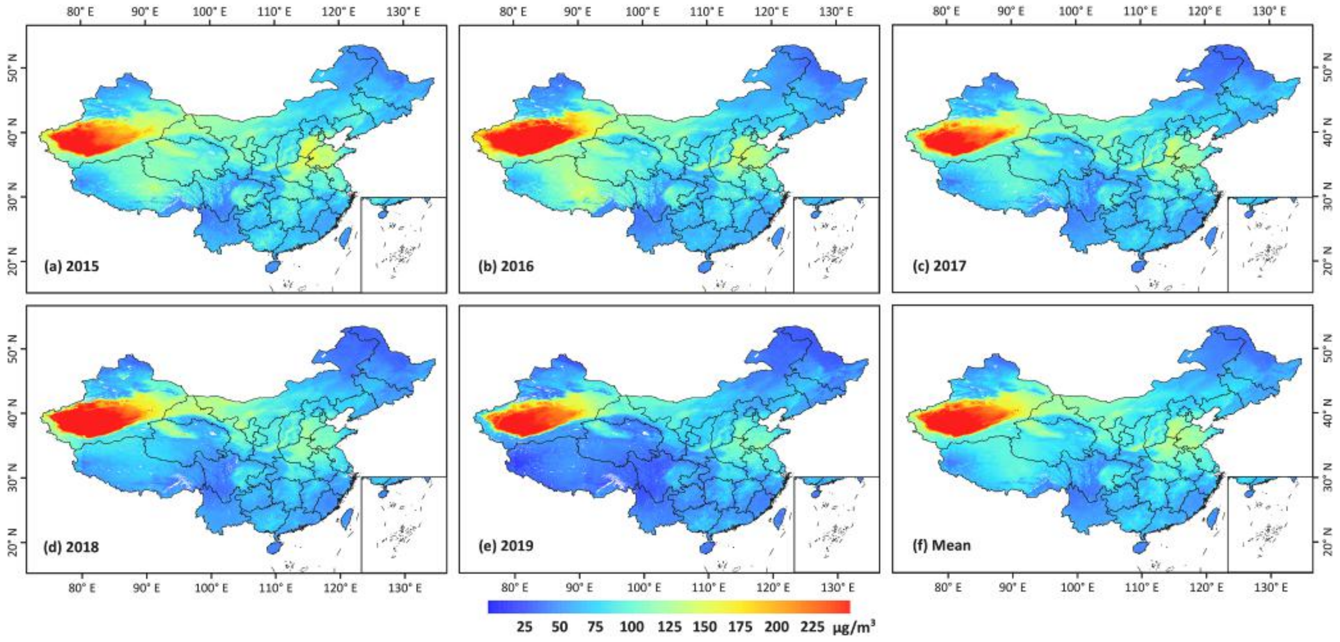
80° E 90° E 100° E 110° E 120° E 130° E
20° N 30° N 40° N 50° N
(a) 2015
(b) 2016
(c) 2017
(d) 2018
(e) 2019
(f) Mean
25 50 75 100 125 150 175 200 225 μg/m³

Fig. 8. Annual (a-e) and multiple-year (f) mean PM_{10} maps (1 km) from 2015 to 2019 across mainland China.

Table 3
Statistics of annual and seasonal mean PM_{10} concentrations (μg/m^3) in China and in each region of interest from 2015 to 2019.

Region	Annual	Spring	Summer	Autumn	Winter
China	85.4 ± 47.6	101.2 ± 65.3	69.4 ± 45.5	79.1 ± 40.0	97.6 ± 44.6
WCH	70.2 ± 21.4	78.8 ± 22.8	46.7 ± 14.2	66.4 ± 21.6	87.2 ± 30.8
ECH	94.8 ± 55.8	115.1 ± 77.8	81.0 ± 53.0	86.9 ± 46.0	104.4 ± 50.3
BTH	93.0 ± 22.7	109.5 ± 19.2	67.5 ± 15.2	86.9 ± 23.1	108.3 ± 35.3
YRD	76.1 ± 15.4	80.7 ± 16.3	52.4 ± 10.7	74.2 ± 16.0	95.0 ± 18.2
PRD	60.0 ± 5.4	64.1 ± 5.9	45.4 ± 6.0	58.5 ± 6.0	68.2 ± 8.4
SCB	68.8 ± 15.2	72.0 ± 14.7	46.5 ± 8.8	62.3 ± 14.7	96.2 ± 25.5

BTH: Beijing-Tianjin-Hebei; ECHN: eastern China; PRD: Pearl River Delta; SCB: Sichuan Basin; WCHN: western China; YRD: Yangtze River Delta.

distributed ground-based monitoring stations, leading to inevitable differences from the actual situation, especially for the first few years. The annual mean PM_{10} concentration is 93.6 ± 43.3, 94.4 ± 52.8, 85.5 ± 44.6, 84.8 ± 56.0, and 69.0 ± 47.2 μg/m^3 for each year from 2015 to 2019, respectively, and the multi-year average is 85.4 ± 47.6 μg/m^3 in China.

Annual PM_{10} concentrations vary across mainland China. Northwest China has extremely high PM_{10} concentrations, especially in the Tarim Basin (e.g., PM_{10} > 250 μg/m^3), because it is the main source area of sand/dust in China. The North China Plain, especially the BTH region, also has high PM_{10} concentrations (average = 93.0 ± 22.7 μg/m^3). This is most likely due to construction-generated emissions or long-range transport of sand and dust (Sun et al., 2006; Chen et al., 2007; Huang, 2010; Liu et al., 2014). By contrast, PM_{10} pollution is generally low in southwestern, northeastern, and southern China (e.g., PM_{10} < 80 μg/m^3), in particular, the PRD region (average = 60.0 ± 5.4 μg/m^3), mainly due to less anthropogenic aerosols or more favorable meteorological conditions for the dispersion of pollution. In general, coarse-mode PM_{10} concentrations have similar spatial patterns but with much higher values as fine-mode $PM_{2.5}$ concentrations in most areas in China, like the Tarim Basin (Wei et al., 2019d).

The STET model can generate almost full-scene PM_{10} maps with high spatial coverage ranging from 93% to 99% across China in different seasons (Fig. 9). However, there are unavoidable missing values in a few parts of southern China in summer and northern China in winter due to abundant clouds and permanent snow/ice. Furthermore, PM_{10} concentrations vary greatly and differently across China in different seasons. PM_{10} concentrations are much higher in spring (average = 101.2 ± 65.3 μg/m^3) and winter (average = 97.6 ± 44.6 μg/m^3), especially in the BTH, PRD, and SCB regions (Table 3). In addition, PM_{10} concentrations are particularly high (i.e., PM_{10} > 200 μg/m^3) in the Taklimakan Desert and in northwestern China throughout the year. This is because there are frequent sandstorms and dust, and these areas are mainly dominated by coarse particles, which can be suspended in the atmosphere for a long time (Ge et al., 2014). By contrast, PM_{10} pollution is lightest in summer (average = 69.4 ± 45.5 μg/m^3) with low PM_{10} concentrations < 70 μg/m^3 in most regions. More precipitation and higher air humidity in the summer are instrumental in diffusing and removing atmospheric pollutants (Li et al., 2017c; Su et al., 2017).

3.2.2. *Temporal variation and trend*

Fig. 10 shows the interannual variations in PM_{10} concentration from 2015 to 2019 in China. PM_{10} pollution had a significant decreasing trend of 5.81 μg/m^3/year (p < 0.001) across China during the five years considered. In general, PM_{10} concentrations showed significant downward trends in most areas, especially in the North China Plain and the Qinghai-Tibet Plateau (i.e., trend < -8 μg/m^3/year, p < 0.05). Therefore, we mainly focus on PM_{10} variations in eastern China, where the mean decreasing trend from 2015 to 2019 was 4.38 μg/m^3/year (p < 0.001). In particular, PM_{10} concentrations decreased significantly (p < 0.001) in three major urban agglomerations in China, with an average annual decline of 6.16 μg/m^3, 3.23 μg/m^3, and 2.75 μg/m^3 for the BTH, YRD, and PRD regions, respectively. The SCB also experienced a noticeable

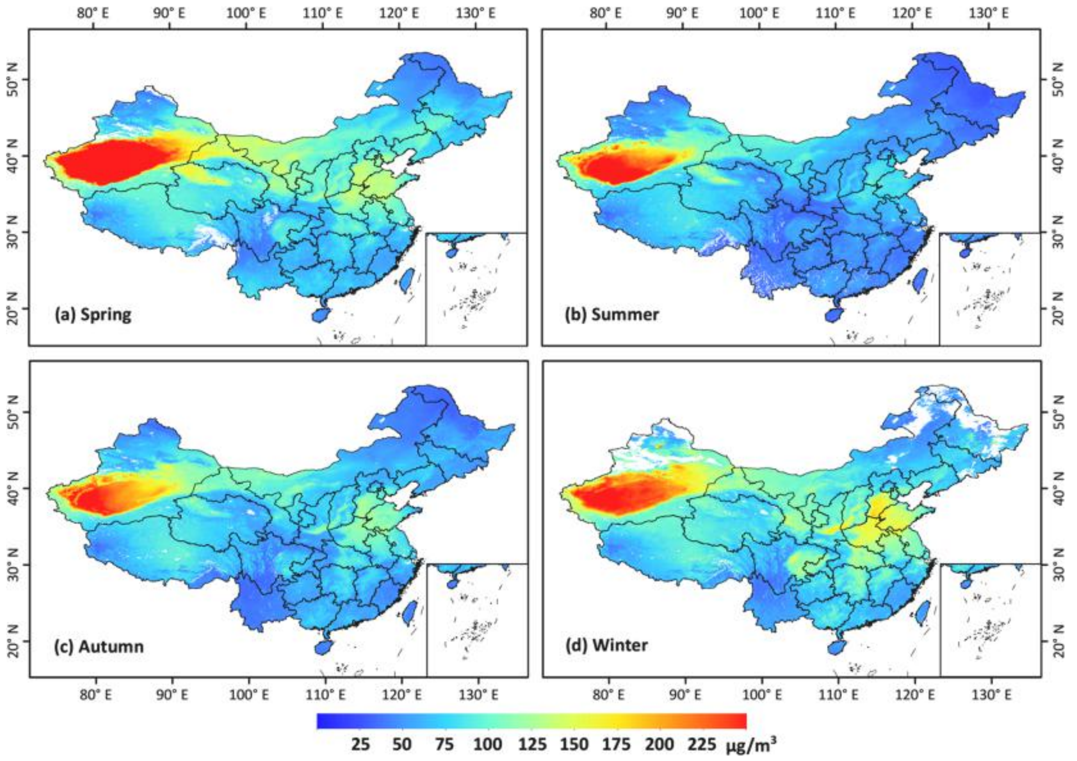

Fig. 9. Seasonal mean PM_{10} maps (1 km) from 2015 to 2019 across mainland China.

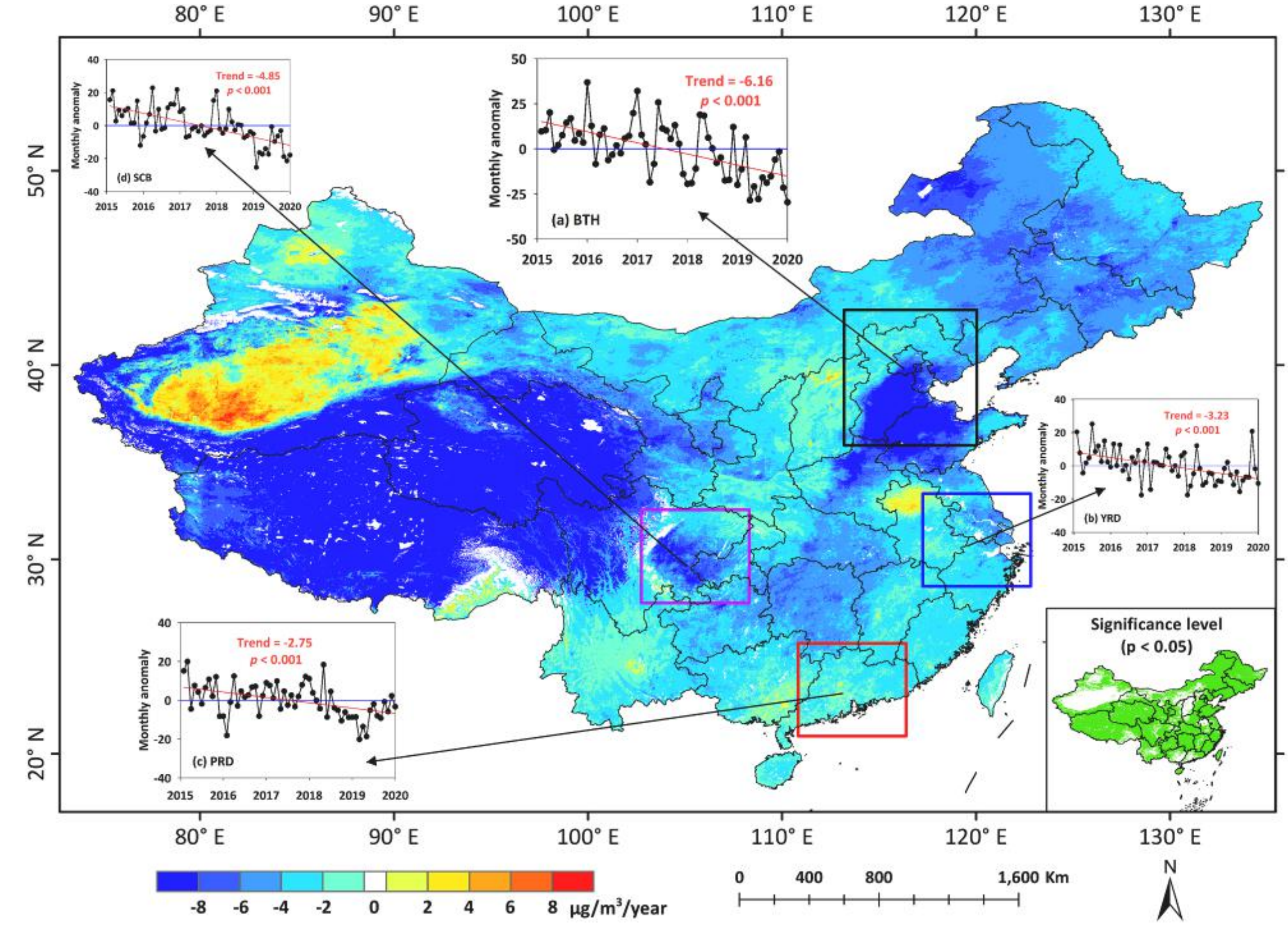

Fig. 10. Spatial distributions of linear trends of monthly PM_{10} concentrations (μg/m^3/year) from 2015 to 2019 across China (background map). Inset figures show time series of the monthly mean anomalies in four regions of interest: (a) the Beijing-Tianjin-Hebei (BTH) region, (b) the Yangtze River Delta (YRD) region, (c) the Pearl River Delta (PRD) region, and (d) the Sichuan Basin (SCB). The red and blue lines represent the linear regression and y = 0 lines, respectively. The green areas in the lower right inset figure represent trends that are significant at the 95% ($p < 0.05$) confidence level. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

decreasing trend of 4.85 μg/m^3/year ($p < 0.001$) in PM_{10} concentration. These decreasing trends are mainly due to a series of environmental protection measures implemented by the Chinese government (Zhang et al., 2019).

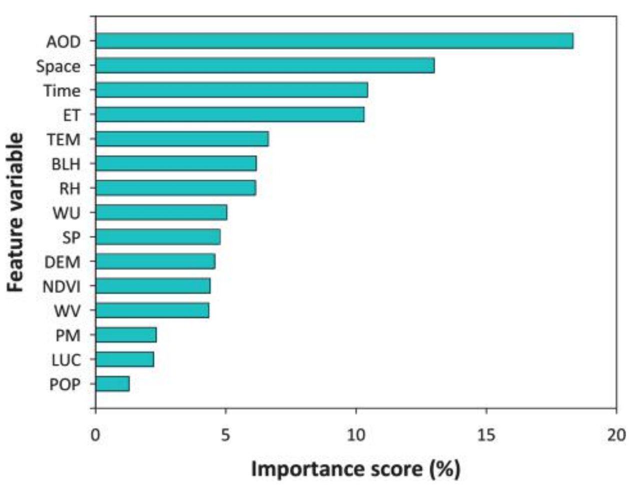

Fig. 11. Sorted importance scores for each feature considered for estimating PM_{10} concentrations in China using the STET model.

3.3. Discussion

3.3.1. Importance of the feature variables

Tree-based ensemble learning methods allow quantitative evaluations of the importance of each input feature variable to the predictor (i.e., PM_{10}), expressed by an importance score. This score is calculated using the Gini index (Jiang et al., 2009; Calle and Urrea, 2011), used to represent the average contributions of features to the model performance in the tree node splitting rather than the physical mechanism. Fig. 11 shows the sorted importance scores for all features used in the STET model to estimate PM_{10} concentrations in China. AOD is clearly the most critical variable, and its importance score reaches 18%. This is followed by space and time, with importance scores of 13% and 10%, respectively. This indicates the importance of spatiotemporal information in improving model performance. In addition, all meteorological variables (especially ET and TEM), DEM, and NDVI have impacts on PM_{10} with varying importance scores ranging from 4% to 10%. The

Table 4
Comparison of the STET model with and without considering spatial distances in PM_{10} estimates in China. Data are from 2019.

Model	Out-of-sample validation			Out-of-station validation		
	R^2	RMSE	MAE	R^2	RMSE	MAE
$STET^{No\ Distance}$	0.856	20.441	12.124	0.827	22.428	13.840
$STET^{Full}$	0.866	19.757	11.530	0.835	21.856	13.453

Table 5
Model performances of the STET model and traditional models using the same input dataset in China, where bold values indicate the optimal value of each column. Data are from 2019.

Model	Spatial resolution	Regression Line			Model Validation	
		Slope	Intercept	R^2	RMSE	MAE
MLR	1 km	0.31	53.07	0.30	24.41	27.95
GAM	1 km	0.36	54.05	0.32	31.31	32.55
GWR	1 km	0.54	37.00	0.50	30.64	25.53
LME	1 km	0.58	32.43	0.57	25.68	21.55
Two-stage	1 km	0.67	25.90	0.65	26.60	20.21
DCT	1 km	0.63	30.96	0.53	37.79	22.54
RF	1 km	0.69	24.96	0.76	25.46	16.19
ERT	1 km	0.70	24.16	0.79	24.33	15.45
STDT	1 km	0.73	23.69	0.69	30.00	17.71
STRF	1 km	0.79	17.12	0.84	20.67	12.73
STET	1 km	**0.81**	**15.26**	**0.87**	**19.76**	**11.53**

DCT: decision tree; ERT: extra-trees; GAM: generalized additive model; GWR: geographically weighted regression; LME: linear mixed-effect; MLR: multiple linear regression; RF: random forest; STDT: space-time decision tree; STET: space-time extremely randomized trees; STRF: space-time random forest.

remaining four variables have relatively low importance scores, mainly due to their low spatiotemporal resolutions. Nevertheless, all selected features contribute to PM_{10} (i.e., importance scores > 1%) and should not be ignored.

We also performed a simple comparison between the performance of our STET model with and without considering spatial distances using the same input data from 2019 in China (Table 4). Incorporating spatial distances into the model leads to improvements in the overall accuracy and spatial prediction ability, demonstrated by increasing CV-R^2 values and decreasing estimation uncertainties, i.e., RMSE and MAE. This indicates that the combination of spatial distances and longitude/latitude information can describe the spatial location of a point in space more accurately, helpful for training a more accurate model (Behrens et al., 2018).

3.3.2. Comparison with traditional models

First, we compared the performance of our STET model with traditional linear regression, statistical regression, and machine learning models using the same input dataset from 2019 in China (Table 5). Results suggest that PM_{10} concentrations derived from the MLR model do not agree well with ground measurements (e.g., slope = 0.31, CV-R^2 = 0.30), showing large estimation uncertainties. The generalized additive model (GAM), GWR, and LME models can improve the overall accuracy of PM_{10} estimates, as shown by better regression lines, increasing CV-R^2 values, and decreasing estimation errors. The two-stage model can also generate more accurate PM_{10} estimates (e.g., slope = 0.67, CV-R^2 = 0.65) with smaller estimation uncertainties (e.g., MAE = 20.21 μg/m^3) by combining two different statistical regression models (i.e., LME and GWR). However, among the three popular tree-based machine learning models, the RF and ERT models perform better in estimating PM_{10} than can the original DCT model because they are ensemble learning approaches developed from the decision tree. They also outperform the statistical regression models considered in Table 5. More importantly, by introducing spatiotemporal information, the performances of the STDT, STRF, and STET models is improved significantly with overall better statistical metrics (CV-R^2 = 0.69–0.87, RMSE = 19–30 μg/m^3, and MAE = 11–18 μg/m^3). In particular, the STET model performs the best with all the best evaluation indicators among all models considered.

3.3.3. Comparison with previous studies

We then compared our results with results from previous related studies on PM_{10} estimations from regional to national scales in China (Table 6). The STET model performs better than the traditional physical model in the BTH region (Wang et al., 2014) and the LME model in the YRD region (Meng et al., 2015). The STET model also outperforms the GWR model (Zhang et al., 2016), the nonlinear exposure-lag-response model (Chen et al., 2018), the GAM model (Chen et al., 2018), the RF model (Chen et al., 2018), and the LUR model (Zhang et al., 2018) across the whole of China. Overall, the STET model results are superior to those from previous studies chiefly because previous studies have mostly used traditional statistical regression and conventional machine learning methods. They either have low data mining abilities or do not consider the spatiotemporal characteristics of air pollution, leading to less robust PM_{10}–AOD relationships. Moreover, the AOD products used in previous studies were generated from either the DT or DB algorithms, with large estimation uncertainties, especially over bright surfaces (Sayer et al., 2014; Wei et al., 2019c, 2020). By contrast, the AOD product we used was generated from the new MAIAC algorithm, which has proven to be much more accurate than the popular DT and DB algorithms, with significantly smaller estimation uncertainties, especially over heterogeneous urban surfaces (Liu et al., 2019b; Mhawish et al., 2019; Wei et al., 2019e).

The spatial resolution of the MAIAC AOD product (1 km) is much higher than the widely used DT/DB products (3–10 km). Accordingly, the spatial resolution of our generated PM_{10} dataset has been improved to 1 km. To stress the advantage of this finer resolution, Fig. 12 shows 1-km PM_{10} estimates in four hot spots that people pay close attention to, i.e., BTH, YRD, PRD, and SCB. The 1-km ChinaHighPM_{10} dataset provides more spatial details and can provide clearer air pollution information in these typical regions than can the 3-km and 10-km estimates reported by others (e.g., Wang et al., 2014; Meng et al., 2015; Zhang et al., 2016; Chen et al., 2018; Zhang et al., 2018), especially at small to medium scales, such as prefecture-level cities. These results further illustrate that the ChinaHighPM_{10} dataset is potentially useful for future urban air pollution and environmental health studies.

Table 6
Model performances from similar previous studies focused on China.

Model	Spatial Resolution	Model Validation			AOD Product	Study region	Literature
		R^2	RMSE	MAE			
Physical	10 km	0.68	–	–	DT	BTH	Wang et al. (2014)
LME	10 km	0.87	19.20	–	DT	YRD	Meng et al. (2015)
LUR	10 km	0.64	–	–	DT	China	Zhang et al. (2018)
NLELR	10 km	0.47	48.96	–	DT/DB	China	Chen et al. (2018)
GAM		0.50	47.40	–	DT/DB		
RF		0.78	31.54	–	DT/DB	–	
GWR	3 km	0.81	–	–	DT	China	Zhang et al. (2016)
STET	1 km	0.89	24.81	15.99	MAIAC	BTH	This study
		0.86	16.40	11.32	MAIAC	YRD	
		0.86	24.28	14.52	MAIAC	China	

DB: Deep Blue; DT: Dark Target; GAM: generalized additive model; GWR: geographically weighted regression; LME: linear mixed-effect; LUR: land-use regression; MAIAC: Multi-Angle Implementation of Atmospheric Correction; NLELR: nonlinear exposure-lag-response; RF: random forest; STET: space-time extremely randomized trees.

4. Summary and conclusion

In view of the low spatial resolutions of current air pollution datasets, a tree-based ensemble learning method, which accounts for spatiotemporal information (i.e., space-time extremely randomized trees, or STET, model), was designed to derive near-surface PM_{10} concentrations using remote sensing technology. For this purpose, and based on the newly released MAIAC 1-km-resolution aerosol product and PM_{10} ground observations, together with auxiliary data, i.e., meteorological, emission, land cover, topography, and human activity data, a daily high-resolution (1 km) and high-quality PM_{10} dataset in China (i.e., ChinaHighPM_{10}) from 2015 to 2019 was produced. The ChinaHighPM_{10} dataset is highly accurate, with high out-of-sample and out-of-station cross-validation R^2 values of 0.86 and 0.82, respectively, and low RMSE values of 24.28 μg/m^3 and 27.07 μg/m^3, respectively, at the national scale. In addition, the STET model outperforms most traditional physical, statistical regression, and machine learning models.

High PM_{10} concentrations are mainly observed in northwestern China and the North China Plain, while low concentrations are always found in southwestern, northeastern, and southern China. However, PM_{10} concentrations have significantly decreased over the past five years in China, with an average trend of −5.81 μg/m^3 ($p < 0.001$), especially in the Beijing-Tianjin-Hebei region (trend = 6.16 μg/m^3, $p < 0.001$). The Tarim Basin, though, still has high PM_{10} concentrations > 250 μg/m^3 because this area is mainly dominated by coarse particles, and sandstorms are common throughout the year. In general, the ChinaHighPM_{10} dataset can provide information useful for future related air pollution studies in small- and medium-scale regions such as urban areas.

CRediT authorship contribution statement

Jing Wei: Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft. **Zhanqing Li:** Funding acquisition, Methodology, Supervision, Writing - review & editing. **Wenhao Xue:** Data curation. **Lin Sun:** Writing - review & editing. **Tianyi Fan:** Writing - review & editing. **Lei Liu:** Writing - review & editing. **Tianning Su:** Writing - review & editing. **Maureen Cribb:** Writing - review & editing.

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.

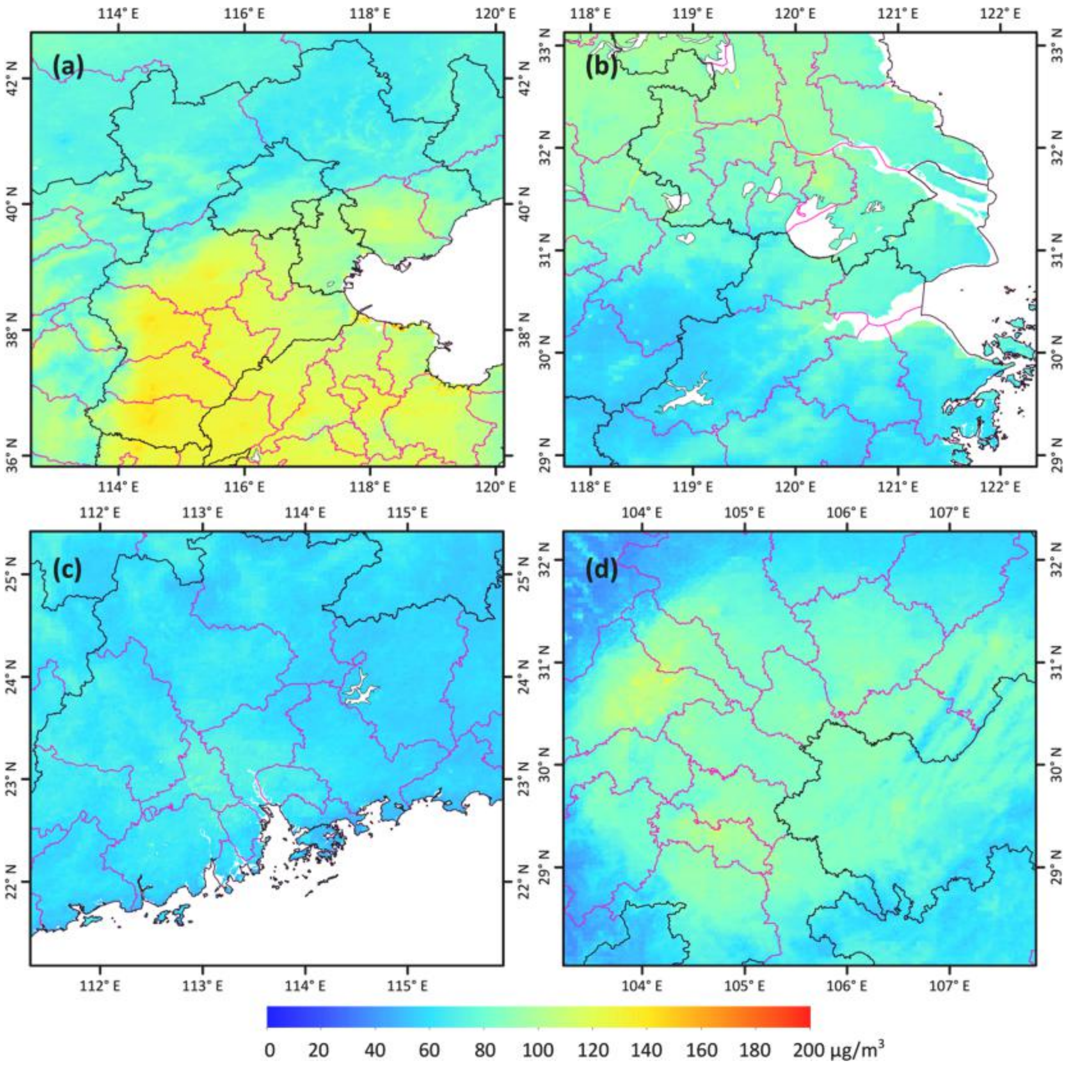

Fig. 12. Multi-year (2015–2019) mean PM_{10} maps (1 km) for (a) the Beijing-Tianjin-Hebei region, (b) the Pearl River Delta region, (c) the Yangtze River Delta region, and (d) the Sichuan Basin in China. The solid black and pink lines represent provincial and city boundaries, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgments

This work was supported by the National Key R&D Program of China (2017YFC1501702), the National Important Project of the Ministry of Science and Technology in China (2017YFC1501404), and the National Natural Science Foundation of China (41705125). The first author is funded by the Shanghai Tongji Gao Tingyao Environmental Science & Technology Development Foundation. We would like to thank Dr. Qiang Zhang at Tsinghua University for providing MEIC emission data for China. The ChinaHighPM_{10} dataset is available at https://weijing-rs.github.io/product.html (released on 15 April 2020).

References

Bartell, S., Longhurst, J., Tjoa, T., Sioutas, C., Delfino, R., 2013. Particulate air pollution, ambulatory heart rate variability, and cardiac arrhythmia in retirement community residents with coronary artery disease. Environ. Health Perspect. 121, 1135–1141.

Behrens et al., Schmidt, K., Rossel, R., Gries, P., Scholten, T., Macmillan, R., 2018. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci., September 2018, 69, 757–770.

Beloconi, A., Kamarianakis, Y., Chrysoulakis, N., 2016. Estimating urban PM_{10} and $PM_{2.5}$ concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data. Remote Sens. Environ. 172, 148–164.

Benas, N., Beloconi, A., Chrysoulakis, N., 2013. Estimation of urban PM_{10} concentration, based on MODIS and MERIS/AATSR synergistic observations. Atmos. Environ. 79, 448–454.

Bi, X., Feng, Y., Wu, J., Wang, Y., Zhu, T., 2007. Source apportionment of PM_{10} in six cities of northern China. Atmos. Environ. 41 (5), 903–912.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Calle, M., Urrea, V., 2011. Letter to the editor: satiability of random forest importance measures. Brief. Bioinform. 12, 86–89.

Chan, C., Yao, X., 2008. Air pollution in mega cities in China. Atmos. Environ. 42 (1), 1–42.

Chen, D., Cheng, S., Liu, L., Chen, T., Guo, X., 2007. An integrated MM5-CMAQ modeling approach for assessing trans-boundary PM_{10} contribution to the host city of 2008 Olympic summer games—Beijing, China. Atmos. Environ. 41 (6), 1237–1250.

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L., et al., 2018. Spatiotemporal patterns of PM_{10} concentrations over China during 2005–2016: a satellite-based estimation using the random forests approach. Environ. Pollut. 242, 605–613.

Choi, Y., Ho, C., Chen, D., Noh, Y., Song, C., 2008. Spectral analysis of weekly variation in PM_{10} mass concentration and meteorological conditions over China. Atmos. Environ. 42 (4), 655–666.

Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), date of access. https://cds.climate.copernicus.eu/cdsapp#!/home.

Dobson, J., Bright, E., Coleman, P., Durfee, R., Worley, B., 2000. A global population database for estimating populations at risk. Photogramm. Eng. Remote Sens. 66 (7).

Emili, E., Popp, C., Petitta, M., Riffler, M., Wunderle, S., Zebisch, M., 2010. PM_{10} remote sensing from geostationary SEVIRI and polar-orbiting MODIS sensors over the complex terrain of the European alpine region. Remote Sens. Environ. 114 (11), 2485–2499.

Franklin, M., Kalashnikova, O., Garay, M.J., 2017. Size-resolved particulate matter concentrations derived from 4.4 km-resolution size-fractionated Multi-angle Imaging SpectroRadiometer (MISR) aerosol optical depth over Southern California. Remote Sens. Environ. 196, 312–323.

Garay, M.J., Witek, M.L., Kahn, R.A., Seidel, F.C., Limbacher, J.A., Bull, M.A., Diner, D. J., Hansen, E.G., Kalashnikova, O.V., Lee, H., Nastan, A.M., Yu, Y., 2020. Introducing the 4.4 km spatial resolution Multi-Angle Imaging SpectroRadiometer (MISR) aerosol product. Atmos. Meas. Tech. 13, 593–628.

Ge, J., Huang, J., Xu, C., Qi, Y., Liu, H., 2014. Characteristics of Taklimakan dust emission and distribution: a satellite and reanalysis field perspective. J. Geophys. Res.: Atmos. 119 (20), 11772–11783.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.

Ghotbi, S., Sotoudeheian, S., Arhami, M., 2016. Estimating urban ground-level PM10 using MODIS 3-km AOD product and meteorological parameters from WRF model. Atmos. Environ., S1352231016304903.

Guo, J., Zhang, X., Che, H., Gong, S., An, X., Cao, C., et al., 2009. Correlation between pm concentrations and aerosol optical depth in eastern China. Atmos. Environ. 43 (37), 5876–5886.

Guo, J., Su, T., Li, Z., Miao, Y., Li, J., Liu, H., Xu, H., Cribb, M., Zhai, P., 2017. Declining frequency of summertime local-scale precipitation over eastern China from 1970 to 2010 and its potential link to aerosols. Geophys. Res. Lett. 44 (11), 5700–5708.

Huang, Q., 2010. An integrated MM5-CAMX modeling approach for assessing PM_{10} contribution from different sources in Beijing, China. J. Environ. Inform. 15 (2), 47–61.

Hsu, N., Lee, J., Sayer, A., Kim, W., Bettenhausen, C., Tsay, S., 2019. VIIRS Deep Blue aerosol products over land: extending the EOS long-term aerosol data records. J. Geophys. Res.: Atmos. 124, 4026–4053.

Jackson, J.M., Liu, H., Laszlo, I., Kondragunta, S., Remer, L.A., Huang, J., Huang, H.C., 2013. Suomi-NPP VIIRS aerosol algorithms and data products. J. Geophys. Res.: Atmos. 118 (22), 12673–12689.

Ji, D., Wang, Y., Wang, L., Chen, L., Liu, Z., 2012. Analysis of heavy pollution episodes in selected cities of Northern China. Atmos. Environ. 50, 338–348.

Jiang, R., Tang, W., Wu, X., Fu, W., 2009. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinf. 10.

Koelemeijer, R., Homan, C., Matthijsen, J., 2006. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. Atmos. Environ. 40, 5304–5315.

Krumbein, W.C., 1959. Trend surface analysis of contour type maps with irregular control-point spacing. J. Geophys. Res. 64, 823–834.

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, 2989–3034.

Li, M., Zhang, Q., Kurokawa, J., Woo, J., He, K., Lu, Z., Ohara, T., Song, Y., Streets, D., Carmichael, G., Cheng, Y., Hong, C., Huo, H., Jiang, X., Kang, S., Liu, F., Su, H., Zheng, B., 2017a. A mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP. Atmos. Chem. Phys. 17, 935–963.

Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017b. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44 (23), 11985–11993.

Li, Z., Zhao, X., Kahn, R., Mishchenko, M., Remer, L., Lee, K.H., et al., 2009. Uncertainties in satellite remote sensing of aerosols and impact on monitoring its long-term trend: a review and perspective. Ann. Geophys. 27 (7), 2755–2770.

Li, Z., Guo, J., Ding, A., Liao, H., Liu, J., Sun, Y., Wang, T., Xue, H., Zhang, H., Zhu, B., 2017c. Aerosol and boundary-layer interactions and impact on air quality. Natl. Sci. Rev. 4, 810–833.

Liu, C., Chen, R., Sera, F., Vicedo-Cabrara, A.M., Guo, Y., et al., 2019a. Ambient particulate air pollution and daily mortality in 652 cities. New Engl. J. Med. 381, 705–715.

Liu, N., Zou, B., Feng, H., Wang, W., Tang, Y., Liang, Y., 2019b. Evaluation and comparison of multiangle implementation of the atmospheric correction algorithm, Dark Target, and Deep Blue aerosol products over China. Atmos. Chem. Phys. 19, 8243–8268.

Liu, Q., Liu, Y., Yin, J., Zhang, M., Zhang, T., 2014. Chemical characteristics and source apportionment of PM_{10} during Asian dust storm and non-dust storm days in Beijing. Atmos. Environ. 91, 85–94.

Lyapustin, A., Wang, Y., Korkin, S., Huang, D., 2018. MODIS Collection 6 MAIAC algorithm. Atmos. Meas. Tech. 11, 5741–5765.

Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. Environ. Sci. Technol. 48 (13), 7436–7444.

Meng, X., Fu, Q., Ma, Z., Chen, L., Zou, B., Zhang, Y., et al., 2015. Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model. Environ. Poll., S0269749115300890.

Mhawish, A., Banerjee, T., Sorek-Hamer, M., Lyapustin, A., Broday, D., Chatfield, R., 2019. Comparison and evaluation of MODIS Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol product over south Asia. Remote Sens. Environ. 224, 12–28.

Nordio, F., Kloog, I., Coull, B., Chudnovsky, A., Grillo, P., Bertazzi, P., Baccarelli, A., Schwartz, J., 2013. Estimating spatiotemporal resolved PM_{10} aerosol mass concentrations using MODIS satellite data and land use regression over Lombardy, Italy. Atmos. Environ. 74, 227–236.

Qu, W., Arimoto, R., Zhang, X., Zhao, C., Wang, Y., Sheng, L., et al., 2010. Spatial distribution and interannual variation of surface PM_{10} concentrations over eighty-six Chinese cities. Atmos. Chem. Phys. 10 (12), 5641–5662.

Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575.

Rohde, R., Muller, R., 2015. Air pollution in China: mapping of concentrations and sources. PLoS ONE 10, e0135749.

Sayer, A., Munchak, L., Hsu, N., Levy, R., Bettenhausen, C., Jeong, M., 2014. MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and "merged" data sets, and usage recommendations. J. Geophys. Res.: Atmos. 119, 24. https://doi.org/10.1002/2014JD022453.

Sotoudeheian, S., Arhami, M., 2014. Estimating ground-level PM_{10} using satellite remote sensing and ground-based meteorological measurements over Tehran. Iranian J. Environ. Health Sci. Eng. 12 (1).

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., De Hoogh, K., De' Donato, F., et al., 2019. Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179.

Su, T., Li, J., Li, C., Lau, A.K.H., Yang, D., Shen, C., 2017. An intercomparison of AOD-converted $PM_{2.5}$ concentrations using different approaches for estimating aerosol vertical distribution. Atmos. Environ. 166, 531–542.

Su, T., Li, Z., Kahn, R., 2018. Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over China: regional pattern and influencing factors. Atmos. Chem. Phys. 18 (21).

Su, T., Laszlo, I., Li, Z., Jing, W., Kalluri, S., 2020. Refining aerosol optical depth retrievals over land by constructing the relationship of spectral surface reflectances through deep learning: application to Himawari-8. Remote Sens. Environ. 251, 112093.

Sun, L., Wei, J., Duan, D., Guo, Y., Yang, D., Jia, C., Mi, X., 2016. Impact of land-use and land-cover change on urban air quality in representative cities of China. J. Atmos. Sol. Terr. Phys. 142, 43–54.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., Hsu, K., 2018. A review of global precipitation data sets: data sources, estimation, and intercomparisons. Rev. Geophys. 56 https://doi.org/10.1002/2017RG000574.

Sun, Y., Zhuang, G., Tang, A., Wang, Y., An, Z., 2006. Chemical characteristics of $PM_{2.5}$ and PM_{10} in haze−fog episodes in Beijing. Environ. Sci. Technol. 40, 10, 3148–3155.

Wang, Z., Chen, L., Tao, J., Liu, Y., Hu, X., Tao, M., 2014. An empirical method of RH correction for satellite estimation of ground-level PM concentrations. Atmos. Environ. 95, 71–81. https://doi.org/10.1016/j.atmosenv.2014.05.030.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019a. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221.

Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., Song, Y., et al., 2020. Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space-time extremely randomized trees. Atmos. Chem. Phys. 20 (6), 3273–3289.

Wei, J., Li, Z., Guo, J., Sun, L., Huang, W., Xue, W., Fan, T., Cribb, M., 2019b. Satellite-derived 1-km-resolution PM_1 concentrations from 2014 to 2018 across China. Environ. Sci. Technol. 53 (22), 13265–13274. https://doi.org/10.1021/acs.est.9b03258.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., Cribb, M., 2021. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Remote Sens. Environ. 252, 112136.

Wei, J., Li, Z., Peng, Y., Sun, L., 2019c. MODIS Collection 6.1 aerosol optical depth products over land and ocean: validation and comparison. Atmos. Environ. 201, 428–440.

Wei, J., Peng, Y., Mahmood, R., Sun, L., Guo, J., 2019d. Intercomparison in spatial distributions and temporal trends derived from multi-source satellite aerosol products. Atmos. Chem. Phys. 19, 7183–7207.

Wei, J., Li, Z., Peng, Y., Sun, L., Yan, X., 2019e. A regionally robust high-spatial-resolution aerosol retrieval algorithm for MODIS images over Eastern China. IEEE Trans. Geosci. Remote Sens. 57 (7), 4748–4757.

Wei, J., Li, Z., Sun, L., Peng, Y., Liu, L., He, L., Qin, W., Cribb, M., 2020. MODIS Collection 6.1 3 km resolution aerosol optical depth product: global evaluation and uncertainty analysis. Atmos. Environ. 240, 117768.

Xu, P., Chen, Y., Ye, X., 2013. Haze, air pollution, and health in China. Lancet 382 (9910), 2067.

Yao, F., Wu, J., Li, W., Peng, J., 2019. A spatially structured adaptive two-stage model for retrieving ground-level $PM_{2.5}$ concentrations from VIIRS AOD in China. ISPRS J. Photogramm. Remote Sens. 151, 263–276.

Yoshida, M., Kikuchi, M., Nagao, T., Murakami, H., Nomaki, T., Higurashi, A., 2018. Common retrieval of aerosol properties for imaging satellite sensors. J. Meteorol. Soc. Jpn 96b, 193–209.

You, W., Zang, Z., Zhang, L., Li, Z., Chen, D., Zhang, G., 2015. 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.

You, W., Zang, Z., Zhang, L., Zhang, M., Pan, X., Li, Y., 2016. A nonlinear model for estimating ground-level PM_{10} concentration in Xi'an using MODIS aerosol optical depth retrieval. Atmos. Res. 168, 169–179.

Zaman, N.A.F.K., Kanniah, K.D., Kaskaoutis, D., 2017. Estimating particulate matter using satellite based aerosol optical depth and meteorological variables in Malaysia. Atmos. Res. 193, 142–162.

Zang, L., Mao, F., Guo, J., Gong, W., Wang, W., Pan, Z., 2018. Estimating hourly PM_1 concentrations from Himawari-8 aerosol optical depth in China. Environ. Pollut. 241, 654–663.

Zhang, Y., Li, Z., 2015. Remote sensing of atmospheric fine particulate matter ($PM_{2.5}$) mass concentration near the ground from satellite observation. Remote Sens. Environ. 160, 252–262.

Zhang, Q., Streets, D., He, K., Klimont, Z., 2007. Major components of China's anthropogenic primary particulate emissions. Environ. Res. Lett., 2, No. 045027.

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., et al., 2019. Drivers of improved PM2.5 air quality in China from 2013 to 2017. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.1907956116.

Zhang, T., Gong, W., Zhu, Z., Sun, K., Huang, Y., Ji, Y., 2016. Semi-physical estimates of national-scale PM_{10} concentrations in China using a satellite-based geographically weighted regression model. Atmosphere 7, 88. https://doi.org/10.3390/atmos7070088.

Zhang, Z., Wang, J., Hart, J.E., Laden, F., Zhao, C., Li, T., et al., 2018. National scale spatiotemporal land-use regression model for $PM_{2.5}$, PM_{10} and NO_2 concentration in China. Atmos. Environ. 192, 48–54.

Zhang, Z., Wu, W., Fan, M., Tao, M., Wei, J., Tan, Y., Wang, Q., 2019. Validation of Himawari-8 Aerosol Optical Depth Retrievals over China. Atmos. Environ. 199, 32–44.