Atmospheric Research 281 (2023) 106481

Contents lists available at ScienceDirect

Atmospheric Research

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

Spatiotemporal high-resolution imputation modeling of aerosol optical depth for investigating its full-coverage variation in China from 2003 to 2020

Qingqing He [a,*], Weihang Wang [a], Yimeng Song [b], Ming Zhang [a], Bo Huang [c]

[a] School of Resource and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China
[b] School of the Environment, Yale University, New Haven, CT, 06511, USA
[c] Department of Geography and Resource Management, The Chinese University of Hsong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Aerosol Optical Depth (AOD)
MAIAC
Full coverage
Long-term trend
Random forest

ABSTRACT

Investigating spatiotemporal variations of atmospheric aerosols is important for climate change and environmental research. Although satellite aerosol optical depth (AOD) retrieved by the MAIAC (Multiangle Implementation of Atmospheric Correct) algorithm provides a unique opportunity to represent global aerosol loading with high spatiotemporal resolution, accurate assessment of long-term aerosol loading countrywide is still challenging due to its non-random missingness. This study aimed to develop an adaptive spatiotemporal high-resolution imputation modeling framework for AOD that incorporates random forest models and multisource data (the simulated AOD, meteorological, and surface condition data) to support full-coverage long- and short-term aerosol studies in China. Aided by the time-stratified approach, the imputation model was constructed for each day, and the MAIAC AOD was used as the target variable. The proposed approach could effectively capture the massive spatiotemporal variability in a large amount of data and deliver full-coverage AODs with high accuracies at a daily timescale (i.e., overall validation R^2 against ground-level AOD measurements of 0.77). We then employed the proposed approach to impute the daily MAIAC retrieved AOD towards complete coverage for China for 2003–2020. Due to the complete coverage, the spatial pattern of monthly/seasonal/yearly mean AOD imputations has better representativeness than that of original MAIAC retrievals. Comparison analysis shows that the monthly/seasonal/yearly aerosol loading over most of China tends to be underestimated by temporal aggregates of original satellite-retrieved AODs. Such underestimation is particularly severe in summer and over the North China Plain (the amount of underestimation >0.2). Consequently, our full-coverage AOD imputations can advance scientific research and environmental management by supporting national and local complete pictures of both short-term episodes and long-term trends in atmospheric aerosols.

1. Introduction

As tiny particles suspended in the atmosphere, aerosols play a critical role in regional and global climate change through direct radiative forcing and indirect processes such as atmospheric circulation and cloud formation (Kaufman et al., 2002; Li et al., 2007). Ground-level aerosol particles, especially particles with an effective diameter ≤ 10 μm (PM_{10}) and fine particles with a diameter ≤ 2.5 μm ($PM_{2.5}$), have been associated with various adverse effects on human health (Brauer et al., 2016; Collaborators, 2018; Pope and Dockery, 2006). With rocketing industrialization and urbanization in China, the increase in air pollutant emissions has led to a substantial increase in atmospheric aerosol particles, making the country one of the global aerosol hotspots (Guo et al., 2011; Mao et al., 2014; Van Donkelaar et al., 2016; Zheng et al., 2021). To remedy this crisis, the Chinese government has rolled out a series of emission control policies and air pollution prevention actions, which have not only improved air quality but also affected the climate as well as particle-related health burden studies (Ma et al., 2019; Wei et al., 2021; Xue et al., 2019a; Yue et al., 2020). Consequently, accurately quantifying the spatiotemporal variation of atmospheric aerosols in China constitutes fundamentals for solving scientific questions and formulating environmental policies (Chudnovsky et al. 2013; He et al., 2021b).

Many studies have explored the spatiotemporal pattern of

* Corresponding author.
E-mail address: qqhe@whut.edu.cn (Q. He).

https://doi.org/10.1016/j.atmosres.2022.106481
Received 29 July 2022; Received in revised form 6 October 2022; Accepted 20 October 2022
Available online 23 October 2022

atmospheric aerosols with ground-level monitoring data and satellite remote sensing data (He et al., 2021a; Li et al., 2017; Song et al., 2014; Tao et al., 2015; Van Donkelaar et al., 2016; Wei et al., 2021). Compared to in-situ AOD measurements (Holben et al., 1998), satellite-derived AOD data can effectively support long-term studies over a larger geographic area. Particularly, the MAIAC (Multiangle Implementation of Atmospheric Correct) aerosol product (Lyapustin et al., 2011; Lyapustin et al., 2018) derived from Terra and Aqua satellites has provided the public with a unique opportunity to explore the fine-scale variations of AOD, due to its high spatiotemporal resolution (daily, 1 km). Thus, this publicly available high-resolution aerosol product has been increasingly used in scientific research of investigating atmospheric aerosols (Abuelgasim et al., 2021) and estimating ground-level $PM_{2.5}$ (Di et al., 2016; He et al., 2021a).

However, most studies based on satellite AOD data have been challenged by the data availability issue. Affected by cloud/snow cover and bright surfaces, satellite-derived AOD usually suffers from 40%–80% of missing data (He et al., 2019; Song et al., 2019; Tao et al., 2015). Although the MAIAC algorithm can make more retrievals than previously popular aerosol products derived by Dark Target (DT) and Deep Blue (DB) retrieval algorithms (Chudnovsky et al., 2013a; Chudnovsky et al., 2013b; Levy et al., 2013; Mei et al., 2019), the non-random missingness of MAIAC AOD is still quite an issue. In China, the MAIAC AOD was missing over 50% (Bi et al., 2019; He et al., 2021a; Liu et al., 2019b). Such missingness makes satellite AOD data discontinuous in space and time, introducing additional uncertainty in AOD-derived $PM_{2.5}$ data (Chen et al., 2020; Song et al., 2019; Xiao et al., 2021; Xue et al., 2019b) as well as downstream applications of $PM_{2.5}$ exposure and health impact (He et al., 2021b; Ma et al., 2016; Xiao et al., 2017). More importantly, the spatiotemporally discontinuous aerosol data cannot explore short-term aerosol events, e.g., investigating the evolution and mechanism of an extreme haze episode, nor is it sufficient to support the short-term epidemiological research. Therefore, it is of the utmost importance to fill AOD gaps to authentically represent the fine-scale gradients and long-term variations of atmospheric aerosols.

Numerous methods have been developed to make up gaps left by satellite AOD. More AOD values can be gained, through improving aerosol retrieval algorithms to make them more appropriate for local atmospheric and geographic conditions (Li et al., 2012). Statistical methods are also helpful for increasing AOD values, by consolidating AOD datasets with various spatiotemporal coverages and resolutions, e.g., the MODIS Aqua/Terra 3-km DT and 10-km DB AOD datasets and the Multi-angle Imaging SpectroRadiometer (MISR) Terra 4.4-km AOD dataset (He et al., 2019; Jinnagara Puttaswamy et al., 2014; Ma et al., 2019; Sogacheva et al., 2018; Wang et al., 2019; Xu et al., 2015; Zhao et al., 2021). Despite improvements, those strategies cannot increase AOD coverage to 100%. Another gap-filling approach is to incorporate surface $PM_{2.5}$ information into multi-step spatial methods (e.g., linear regression + interpolation), as developed by Goldberg et al. (2019), Lv et al. (2017), and Kloog et al. (2012), which could improve the data integrity to nearly 100%. However, it is not suitable to gap-fill long-term data for China because the ground-level $PM_{2.5}$ monitoring data required by the modeling approach was rarely available in China in historical years (i.e., prior to 2013). In addition, previous studies that adopted statistical models or spatial methods show a limited capability to capture the variability in the data, especially for a large heterogeneous area, resulting in degraded accuracies for imputed AOD values; the validated R^2 values against ground-level aerosol measurements ranged between 0.18 and 0.52 (Goldberg et al., 2019; Lv et al., 2017; Xiao et al., 2017).

Compared to traditional statistical approaches, data-driven machine learning algorithms can handle a large number of predictors and mine helpful information from vast amounts of input data with tremendous variability (Bai et al., 2022; Bi et al., 2019; Chen et al., 2020; Huang et al., 2021; Lin et al., 2016). Combined with external information from numerical simulation models (e.g., coarse-resolution AOD data), recent studies employed ensemble learning/deep learning approaches (Bai et al., 2022; Bi et al., 2019; Huang et al., 2021; Jiang et al., 2021; Kianian et al., 2021; Li, 2020, 2021; Li et al., 2020; Lops et al., 2021; Pu and Yoo, 2021; Zhao et al., 2019) to impute satellite AOD gaps and thereby achieved full-coverage datasets with high spatiotemporal resolution. However, most studies involved in filling AOD gaps aimed to estimate ground-level $PM_{2.5}$ concentrations, where the imputed AOD was used as an intermediate for subsequent $PM_{2.5}$ modeling rather than a final product for assessing long-term variations of atmospheric aerosols. The effect of satellite AOD gaps on the spatiotemporal representativeness of atmospheric aerosols, especially on longer timescales (e.g., monthly and seasonal), has yet been well explored.

Therefore, this study aimed to develop an adaptive spatiotemporal high-resolution imputation modeling framework relying on ensemble learning to generate full-coverage AOD time series and comprehensively investigate spatiotemporal variations of atmospheric aerosols. MAIAC AOD with the high spatiotemporal resolution was used as the target variable, and the publicly available multisource data reflecting coarse-resolution gradients of AOD, meteorology, and geographical conditions were used as inputs for imputation modeling. The outputs of AOD imputations were validated against MAIAC original AOD retrievals and ground-level AOD measurements. Then, we generated a full-coverage, high-spatiotemporal-resolution (i.e., 1 km, daily) AOD dataset for China from 2003 to 2020. We also carried out a spatiotemporal analysis of full-coverage AOD across China. The uncertainty due to the MAIAC AOD absence and the difference in spatial representativeness between MAIAC and imputed AOD were discussed.

2. Data and methods

2.1. Data

2.1.1. MAIAC AOD

The MAIAC retrieval algorithm (Lyapustin et al., 2018) derives daily aerosol properties at a spatial resolution of 1 km from MODIS instruments aboard the Terra (equatorially crossing at ~10:30 am local time) and Aqua (crossover at 1:30 pm local time) satellites, respectively. The MAIAC aerosol retrieval product over China has been rigorously validated and achieved a similar or even better accuracy ($R^2 > 0.75$) against ground-level aerosol observations with better spatial resolultion, compared to previously widely-used coarse-resolution aerosol products, e.g., MODIS 10-km and 3-km Dark Target AOD datasets (Levy et al., 2013; Liu et al., 2019b; Remer et al., 2013; Zhang et al., 2019). We downloaded both Terra and Aqua MAIAC retrieved AOD data covering China for 18 years from January 2003 to December 2020 from the NASA Earth Data portal (https://earthdata.nasa.gov/earth-observation-data). Gridded AOD values at 550 nm were extracted, and those marked as cloud contamination or covered by water or snow were excluded. Because we aim to yield complete AOD data daily, per-pixel daily averages of Terra and Aqua MAIAC AOD retrievals were computed (referring to (He et al., 2021a)) and used as a dependent variable for the spatiotemporal imputation.

Previous studies have shown the monthly AOD pattern to be a promising predictor to represent the spatial and long-term trend of AOD variability for the high-resolution imputation (Li et al., 2020). Therefore, the monthly mean MAIAC AOD was derived from the per-pixel daily averages of Terra and Aqua AODs. We counted the number of days with available AOD per pixel for each month. The monthly average was only computed for the pixel where days of valid AOD were >50% within a calendar month and not calculated for that without enough AODs in a month.

2.1.2. AERONET AOD

Since AERONET provides aerosol observations with extremely low uncertainty, the AERONET ground-level AOD data were used as "ground truth" to evaluate the accuracy of our imputed AOD (Holben et al., 1998). Version 3 Level 2.0 AERONET AOD data with quality assurance

and cloud screening from 52 sites during 2003–2019 (Fig. 1) across China were acquired from Goddard Space Flight Center (https://aeronet.gsfc.nasa.gov/). Due to the lack of AOD at 550 nm, the AERONET data were logarithmically interpolated to 550 nm from the closest available wavelengths (440 nm and 500 nm) using the Angstrom exponent reported by AERONET (He et al., 2019). Corresponding to the temporal interval of our daily imputed AOD, the 5-min instantaneous AERONET AOD measurements within one calendar day were averaged. Spatially, satellite/imputed AODs within 3 × 3 pixels of the AERONET site's location were averaged to match the gridded and AERONET AOD data.

2.1.3. Reanalysis data

ECMWF (European Centre for Medium-Range Weather Forecasts) atmospheric Composition Reanalysis 4 (EAC4, https://www.ecmwf.int/en/forecasts/dataset/cams-global-reanalysis) is a global atmospheric composition reanalysis data product, which combines multiple remote sensing and ground-level observation data into a globally spatiotemporal continuous dataset using a numerical model of the atmosphere (Inness et al., 2019). It provides temporal high-resolution (3-h) total and species AOD data with a coarse spatial resolution of 0.75° × 0.75° at various wavelengths (e.g., 550 nm, 865 nm) for a long period from 2003 onwards. Recently, the EAC4 AOD has been increasingly used in atmospheric studies (Schneider et al., 2020; Stafoggia et al., 2019). We validated the EAC4 total AOD at 550 nm against AERONET observations, and comparison results in Fig. S1 of the supporting document show that it is generally reliable with an overall r of 0.80. With complete spatial coverage, high temporal resolution, and reliable data quality for a very long period, EAC4 aerosol data were indispensable for our high spatiotemporal AOD imputation. We first considered EAC4 total AOD and AOD species (including black carbon, dust, organic matter, sea salt, and sulfate AOD) for imputation modeling; however, the sensitivity analysis shows that including component AOD cannot significantly improve the model performance but increase the computational cost. Thus, the 3-h EAC4 total AOD at 550 nm for the 2003–2020 study period was averaged to the daily timescale and used for the following imputation modeling. For simplicity, we refer to EAC4 total AOD at 550 nm as EAC4 AOD hereafter.

Meteorological and cloud cover conditions are integral to estimating high spatiotemporal AOD because they are highly related to the formation and dispersion of aerosol particles in the atmosphere (Xiao et al., 2017; Yu et al., 2015). Meteorological and cloud cover variables in the study period were acquired from ERA5 (https://cds.climate.copernicus.eu/cdsapp#!/home), a global climate and weather dataset by ECMWF reanalysis offering temporal high-resolution meteorological data with 0.25° × 0.25° spatial resolution from 1979 to the present. Hourly total column water vapor (kg/m^2), 10 m u- and v-component of wind (m/s), 2 m temperature (K), and total precipitation (m), and low/medium/high/total cloud cover(unitless) were obtained and averaged to daily values.

2.1.4. Surface conditions and dummy variables

Surface conditions have been proven to play critical roles in portraying the spatial variation of AOD (He et al., 2019); therefore, we incorporated elevation and NDVI (Normalized Difference Vegetation Index) in this study for our spatiotemporal high-resolution AOD imputation modeling. As a measure of terrain, elevation with a spatial resolution of 30 m was acquired from ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer, https://asterweb.jpl.nasa.gov/gdem.asp) global digital elevation model. Monthly NDVI data were extracted from the MODIS/Aqua Vegetation Indices L3 Global 1-km SIN Grid product (MYD13A3).

To account for the spatial variations in AOD, we also derived the coordinates of each 1-km grid cell centroid as a spatial predictor. In addition, a temporal dummy variable was included to explain the daily variations.

2.1.5. Data integration

A 1-km grid was created based on the 1-km MAIAC AOD for the study area. To spatially match the size of the created grid, those reanalysis data (i.e., EAC4 AOD and ERA5 meteorological and cloud cover data) with various spatial resolutions were converted to the 1-km gridded values using inverse distance weighted interpolation, and the 30-m

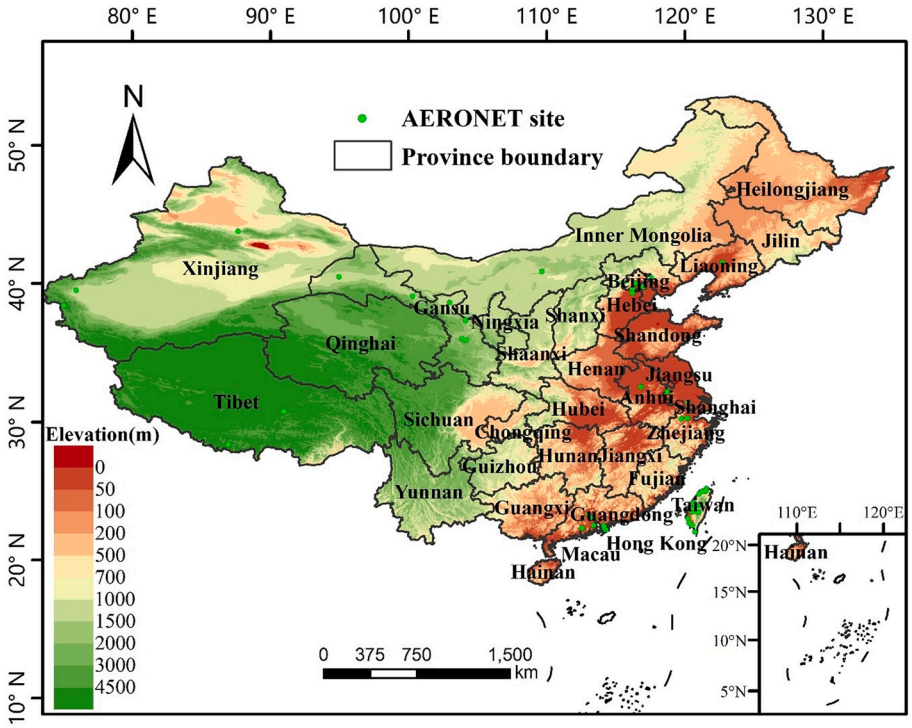

Fig. 1. The study area for this work, presenting the distribution of AERONET sites and terrain.

elevation data was aggregated by calculating the area-weighted mean value within each 1-km grid cell. Then, all the input variables were integrated into the 1-km grid for the spatiotemporal imputation.

2.2. Methods

2.2.1. Random forest

As one machine learning algorithm, random forest (Breiman, 2001) has excellent strength in constructing flexible (linear or non-linear) relationships between response and explanatory variables; with a little configuration, it shows outstanding ability to tackle massive predictor data (Huang et al., 2021; Jiang et al., 2021). Recently, it has been emerging in imputing massive non-randomly missing AOD values in recent years (Chen et al., 2019; Xiao et al., 2021). Thus, this study employed a random forest algorithm to impute daily per-pixel AOD values for China, considering the imputation precision and gigantic data in this work.

Random forest first constructs a set of decision trees, then trains each tree in the forest using bootstrap samples drawn from original input samples, and finally makes predictions as model output by aggregating the predictions of all trees. It uses a randomly selected subset of predictors at each decision split to develop each decision tree. During implementation, random forest usually has two core hyperparameters that impact model performance and need to be tuned: the number of decision trees (N_{tree}) and the number of predictors for each decision split (N_{pts}).

2.2.2. Model development

For each day from 2003 to 2020, we developed a random forest model with spatial and temporal characteristics to impute gaps left by the MAIAC retrieval algorithm. However, the average number of samples for each modeling day remains at 3.526 million over the entire study area, which is too huge to develop each imputation model with an acceptable computational cost. Thus, we adopted an adaptive framework to optimize the two hyperparameters (i.e., N_{tree} and N_{pts}) in each random forest model to obtain satisfactory model performance while maintaining acceptable computational costs. The detailed adaptive framework for model development to accurately and efficiently gap-fill AOD is as follows.

(1) Search hyperparameters on coarse-resolution data. Based on the grid size of EAC4 AOD, we upscaled the other predictors to a coarse-resolution scale and used the coarse-resolution sample data during the same months of the study period to obtain each model's initial set of hyperparameters. That is, imputation models during the same months of the study period shared the same initial set of hyperparameters. The N_{tree} and N_{pts} hyperparameters were tuned via Bayesian optimization and a cross-validated approach.
(2) Randomly divide each modeling sample dataset (samples with high spatial resolution) into training and testing folders. 78.69% of samples were used for model development (including model training and hyperparameters adjusting), and the remaining sample data were held out and used for independent validation.
(3) Train each model and adjust hyperparameters. Based on the model setting obtained in step (1), each random forest model was first trained on 5% of samples or at least 150,000 1 km × 1 km grid cells, whichever is larger. Here, we also examined model performance with 10% of samples or at least 300,000 1 km × 1 km grid cells and found that the increase in the sample dataset would not significantly improve the model performance but remarkably increase the computation cost. The initial samples for each daily model were randomly selected from the training folder. Then, the imputations were made and compared with samples from the hold-out folder. If the validation R^2 for a specific day exceeded 0.80, the model would be used to make estimates as output; otherwise, more samples (~10,000 grid cells) were randomly selected from the training folder and were included to search for new optimal hyperparameters for this modeling day. Such a process would be repeated until the daily validation $R^2 > 0.80$. Fig. 2 demonstrates the heuristic workflow of our proposed spatiotemporal high-resolution imputation modeling.

A time-stratified sampling method was also employed to help each imputation model capture short-term variations in AOD, which trained the random forest model on the grid of 1 km × 1 km with three rolling-day samples and the middle day as the target day. A temporal dummy variable, described as [1,2,3], was included in each model to identify the three rolling days. We also considered using five/seven rolling-day samples with the third/fifth day as the target to develop imputation models. However, sensitivity analyses indicate that including samples from a broader temporal window did not significantly improve model performance but did increase the modeling time. Thus, we adopted three rolling-day samples in the final model development.

In addition to simulated AOD, meteorological parameters, and surface conditions, we also introduced geolocation and temporal indicators in each imputation model. As mentioned in Section 2.1.1, we only calculated monthly MAIAC AOD averages for grid cells where at least half a month had MAIAC AOD retrievals. Therefore, two separate imputation models for grid cells with and without valid monthly AOD averages were developed for each day, denoted as mAOD and non-mAOD models. In this manner, the mAOD models contain 12 variables: EAC4 AOD, total column water vapor, u-wind, v-wind, temperature, total cloud cover, NDVI, elevation, longitude, latitude, time index, and monthly MAIAC AOD, while the non-mAOD models only include the first 11 variables. Combining the mAOD and non-mAOD models can be conceptualized as constructing a binary tree model where we initially enforced a manual split according to the data availability of monthly MAIAC AOD.

After validation, imputations of AOD daily values from 2003 to 2020 over grid cells, where the MAIAC AOD retrievals were missing, were generated by our proposed method. Then, we obtained the final daily full-coverage AOD dataset by fusing the MAIAC original daily retrievals and resultant imputations (referring to fused AOD hereafter). The fused AOD data were further used for spatiotemporal analyses in Section 3.2. The final fused AOD dataset is available at https://dataverse.harvard.edu/dataverse/atmospheric_data_by_WHUT.

2.2.3. Model evaluation

To comprehensively evaluate the reliability of the proposed imputation modeling method, we compared our imputed AOD values with MAIAC original retrievals and AERONET measurements. Here, we referred to the AOD estimated by the proposed model as imputed (or predicted) AOD to distinguish the AOD derived by the MAIAC retrieval algorithm (i.e., MAIAC retrieved AOD). Specifically, we conducted the following evaluations:

(1) MAIAC retrieved AOD vs. our imputed AOD. In this validation, we held out approximately 21.31% of the daily MAIAC retrievals from model development and used them as hold-out validation to individually examine the model output's accuracy. We also compared model outputs with MAIAC retrieved AODs at grid cells with available MAIAC retrievals. The details are shown in Section 3.2.1.
(2) Imputed AOD vs. AERONET AOD. This validation was split into two parts: one excluding grid cells with available MAIAC retrievals, and one only for grid cells with MAIAC AOD retrievals. Here, AERONET AOD data were not employed for imputation modeling and spatiotemporally collocated with our imputed AOD values, as described in Section 2.1.2. The comparison results are detailed in Section 3.2.2.

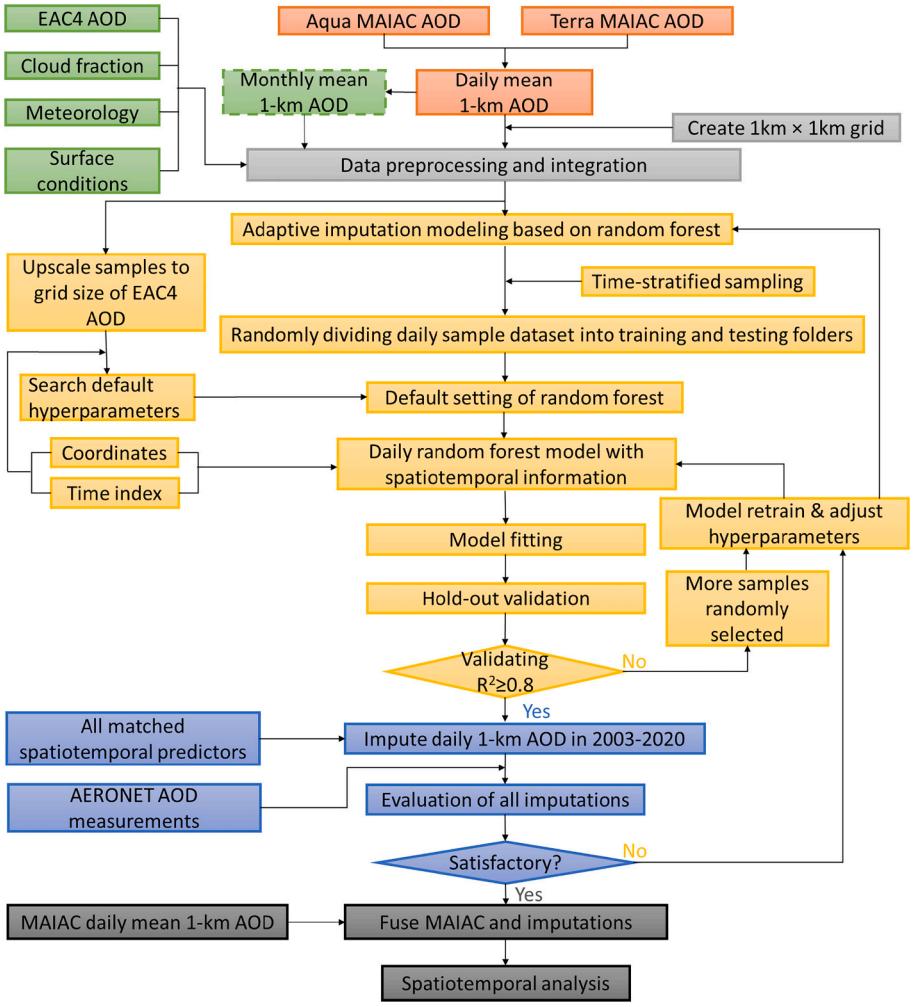

Fig. 2. The heuristic framework of adaptive daily 1-km AOD imputation modeling.

(3) MAIAC retrieved AOD vs. AERONET AOD. This validation was used as a reference to evaluate the accuracy of our imputations.

The Pearson's correlation coefficient r, linearly regressed R^2, and root mean square error (RMSE) were employed to metric the imputation accuracy.

3. Results

3.1. Coverage and descriptive statistics of MAIAC AOD and predictors

The per-pixel data availability of daily mean MAIAC AOD is presented for the entire study period of 2003–2020 (Fig. 3a). The

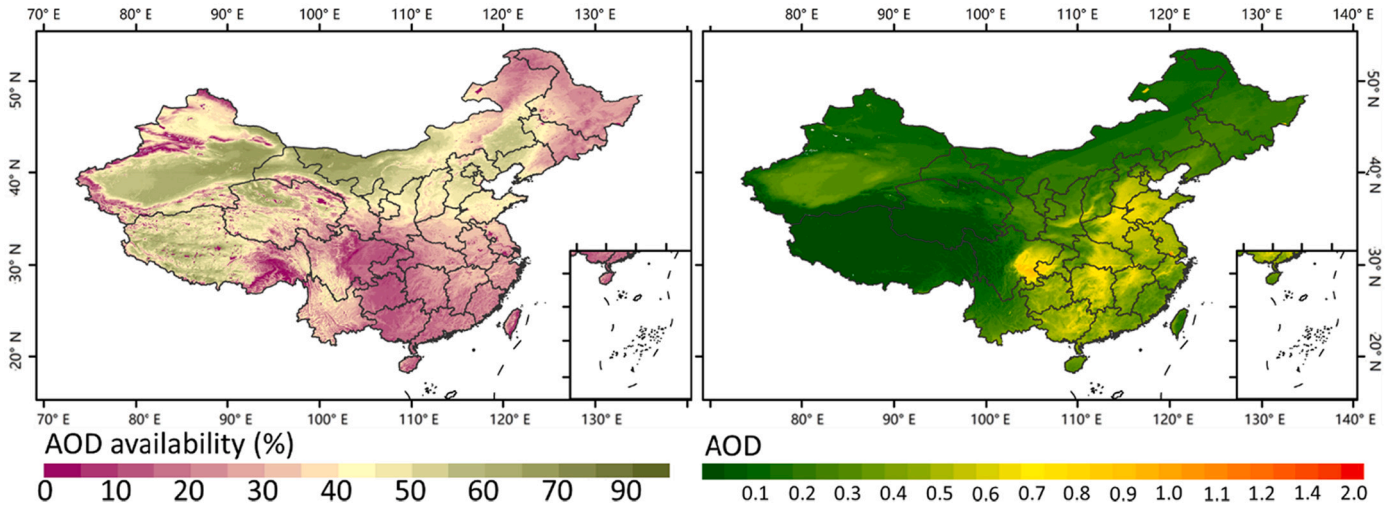

Fig. 3. Spatial distribution of (a) proportion of daily MAIAC AOD availability and (b) MAIAC AOD across China from 2003 to 2020.

proportion of daily MAIAC retrievals is, on average, at 36% over the study period. Spatially, southern China, where most pixels show at most 30% completeness, has lower valid data than northern areas, with the proportion of available retrievals exceeding 40% for most pixels. Fig. S2 demonstrates significant seasonal variations in data availability. The overall proportion of MAIAC retrievals reached its minimum in summer, with a value of 31%, and autumn had the highest availability (45%). Specifically, the mean proportions of per-pixel MAIAC AOD availability were extremely low over eastern Tibet in all seasons and over northeastern and northwestern China in winter, during which the proportions were <10%; it was also very low in Sichuan Basin (SCB) in the whole years, with the per-pixel proportion between 10%–20%. The high reflectance (Lyapustin et al., 2018) and cloud cover are important reasons for the missingness of MAIAC AOD. According to the reanalysis data, the mean cloud fraction value at pixels with the MAIAC retrievals was 27%, while 67% for MAIAC unavailable pixels.

Fig. 3b shows the spatial distribution of multi-year mean MAIAC AOD retrievals from 2003 to 2020 over China. The national 18-year mean MAIAC AOD value was 0.24 with a standard deviation (std) of 0.18. Spatially, the mean AOD was higher in SCB, North China Plain (NCP), and the Hubei-Hunan-Guangxi region, with most pixels having values >0.5, than in other areas. Fig. S3 indicates that the overall AOD observation was higher in spring (0.30) and summer (0.26) than that in winter (0.22) and autumn (0.21), corresponding to previous findings by satellite AOD-related studies (He et al., 2019; Li, 2021). In addition, Table 1 summarizes the descriptive statistics of MAIAC AOD retrievals and other explanatory variables involved in the imputation modeling. After using the time-stratified sampling method in the imputation modeling, the total number of samples is up to 69 billion, which indicates a massive computation cost for the model development.

3.2. Validation results of AOD imputations

3.2.1. Imputation model performance

For the days between 2003 and 2020, we trained 6538 daily random forest models in total. Imputed AOD values from models with and without monthly MAIAC AOD (mAOD model and non-mAOD model) and combined imputations were separately compared with MAIAC AOD retrievals to quantify daily model performance. Time series of daily correlation coefficients between imputed and MAIAC retrieved AOD based on hold-out validation results are presented in Fig. 4. Table S1 summarizes the statistics of hold-out validation results. Overall, our daily imputations generated by combined models show good agreement with MAIAC original AOD retrievals (Fig. S4), with a very high mean validation correlation coefficient r (0.988) and very low mean RMSE (0.036) (Fig. 4). The hold-out validation results also demonstrate that the mAOD and non-mAOD imputation models had similar model performance, with very close r (0.986 vs. 0.987) and RMSE (0.031 for full vs. 0.038 for non-full model) values on average. These comparison results indicate that with the constraint of daily MAIAC AOD retrievals and the assistance of multisource data, imputed AOD values of both models achieved good precision on each modeling day.

Fig. S5 presents the time series of daily correlations between imputed and MAIAC-retrieved AODs based on model-training results, showing that our daily imputation models performed very well throughout the study timespan, with overall r and RMSE values of 0.988 and 0.036 for combined imputations. Comparing Fig. 4 and Fig. S5, daily imputation models did not show apparent differences in statistical metrics between the model training and testing stages. Thus, daily imputation models were slightly overfitted and can provide reliable estimates for pixels where MAIAC AOD retrievals are missing.

3.2.2. Validation results with AERONET data

To further explore the uncertainty of our imputed AOD, we conducted a comparative analysis with AERONET AOD measurements. Here we evaluated daily AOD collocations between AERONET and imputed values at pixels with (Fig. 5b, f) and without (Fig. 5c, g) available MAIAC retrievals, respectively. Fig. 5 (a,e) shows the combined results of mAOD and non-mAOD models. The accuracy based on daily collocations between AERONET and MAIAC-retrieved AODs was also calculated and used as a reference (Fig. 5d, h). In total, we obtained 27,072 daily collocations from combined imputations for validation. The overall R^2 value of daily AERONET-imputed collocations is 0.77 (Fig. 5a), which is very close to the accuracy of the cutting-edge aerosol dataset, MAIAC

Table 1
Sample size and descriptive statistics of sample dataset for China, 2003–2020.

Group	Variable	Mean (std*)				
		All	Spring	Summer	Autumn	Winter
Data volume	No. of samples for each day (million)	3.526 (1.069)	3.406 (0.881)	2.981 (0.687)	4.380 (1.045)	3.343 (1.081)
Target	MAIAC daily AOD	0.230 (0.097)	0.302 (0.120)	0.237 (0.082)	0.188 (0.064)	0.190 (0.059)
	EAC4 AOD	0.262 (0.109)	0.359 (0.110)	0.283 (0.092)	0.202 (0.065)	0.199 (0.073)
Predictor	Temperature (K)	281.100 (10.133)	282.392 (5.567)	293.289 (2.039)	280.248 (6.606)	268.076 (3.627)
	Total cloud cover	0.326 (0.091)	0.354 (0.077)	0.384 (0.077)	0.276 (0.075)	0.289 (0.089)
	Total column water vapor (kg/m2)	10.156 (6.289)	7.920 (2.783)	18.980 (3.991)	9.311 (3.749)	4.229 (1.199)
	U-wind (m/s)	0.595 (0.690)	0.789 (0.700)	0.047 (0.581)	0.581 (0.579)	0.974 (0.509)
	V-wind (m/s)	0.106 (0.557)	0.047 (0.659)	0.028 (0.511)	0.115 (0.523)	0.240 (0.494)
	Elevation (m)	1868.379 (402.418)	1739.038 (342.204)	1721.267 (303.187)	1837.070 (352.520)	2185.283 (421.866)
	NDVI	0.278 (0.082)	0.229 (0.054)	0.365 (0.066)	0.295 (0.064)	0.220 (0.042)
	Longitude (°)	101.382 (3.291)	102.073 (3.473)	101.477 (3.244)	101.850 (3.070)	100.094 (2.990)
	Latitude (°)	37.477 (2.156)	37.681 (1.729)	39.125 (1.390)	37.947 (1.760)	35.090 (1.404)
	MAIAC monthly AOD	0.227 (0.073)	0.298 (0.073)	0.237 (0.053)	0.186 (0.042)	0.186 (0.050)

* std: standard deviation.

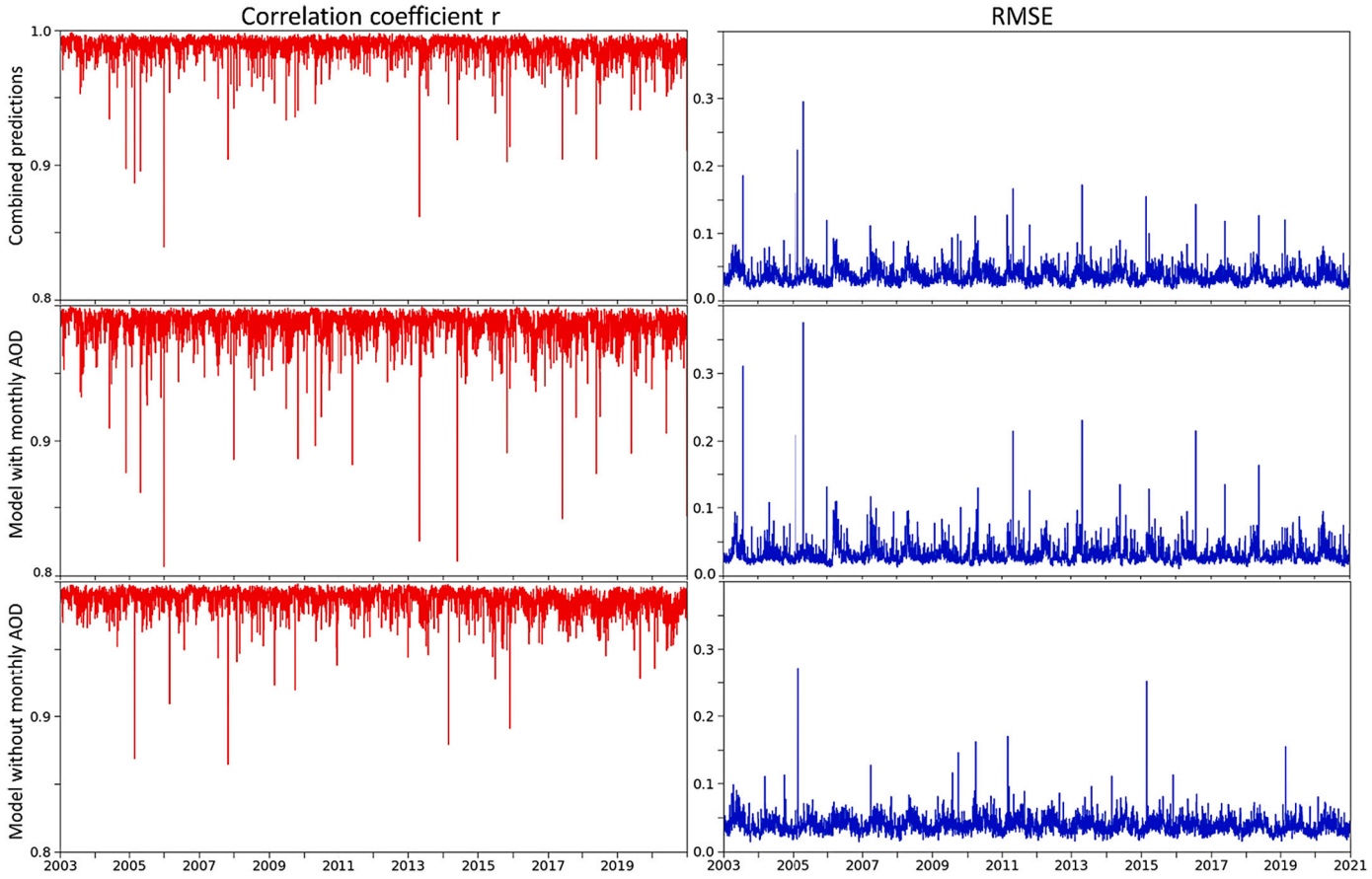

Fig. 4. Time series of daily correlation coefficient r (left panel) and RMSE (right panel) values between imputed and MAIAC-retrieved AOD based on hold-out validation results from 2003 to 2020.

original AOD retrievals over China ($R^2 = 0.82$ in Fig. 5d). The histogram plot shows deviations between AERONET and imputed values follow a normal distribution with a mean (0.006) approaching to 0 (Fig. 5e). The slight degradation of R^2 from 0.82 to 0.77 was attributed to the decrease in imputations where MAIAC cannot yield retrievals. Likewise, Li (2021) also shows a reduced accuracy for imputed AOD at pixels without MAIAC retrievals compared with those combined imputations (validated R^2 of 0.68 vs. 0.78). It is acceptable that a reduced R^2 value exists in areas discarded by the MAIAC algorithm since those areas are usually too complicated (e.g., heavy clouds covering) to make a reliable aerosol retrieval. Despite of ~13% decrease in R^2, our imputations for pixels where MAIAC AOD are missing ($R^2 = 0.70$) are significantly superior to the model input, the original EAC4 AOD ($R^2 = 0.60$ in Fig. S6).

For RMSE, the value based on combined imputations (RMSE = 0.25 in Fig. 5a), especially at MAIAC-unavailable areas (RMSE = 0.34 in Fig. 5c), was higher than that at MAIAC-available areas (RMSE = 0.16 in Fig. 5d). This increase is most likely caused by the larger amount of high AOD values involved in the imputations, which are inclined not to be retrieved by the MAIAC algorithm. The mean values of AERONET measurements for daily collocations where the MAIAC retrievals were available and unavailable were 0.34 and 0.66, respectively. This suggests that aerosol loading is higher in areas where the MAIAC algorithm does not make a retrieval.

We examined the statistics of our imputed vs. AERONET-observed AOD collocations for individual sites to show the spatial uncertainty of the AOD imputations. Fig. S7(a) demonstrates that the overall site-based R^2 value was 0.62 against AERONET AOD, and 25 out of 44 AERONET sites had better agreements, with $R^2 > 0.65$. The imputation models performed better in northern China, exceeding 70% of sites in this region with $R^2 > 0.65$. By contrast, ~50% of sites in southern China possess $R^2 < 0.60$. The spatial difference in accuracy presented by our imputations is mainly related to the spatial variation in the accuracy of MAIAC AOD retrievals (Liu et al., 2019b; Zhang et al., 2019). Comparing Fig. S7 (a) and (d), it is clear that the site accuracy of our AOD imputations shares a similar spatial pattern with MAIAC retrievals across the study region. Fig. S8 presents the monthly comparison results from 2003 to 2020, indicating that the deviations between our imputed and AERONET-observed AODs are consistently small from January to December, with $R^2 > 0.70$ and RMSE<0.35 for all months, and correspond to the seasonal pattern of MAIAC AOD retrievals (Fig. S8 (a)).

Thus, as presented above, the imputed AOD values by our proposed method achieved higher quality, which can be used to represent spatiotemporal variations of atmospheric aerosol loading in China. Then, we obtained the final fused 1-km AOD by consolidating imputed AODs at MAIAC-unavailable grid cells and MAIAC AOD retrievals. Validation results with AERONET measurements show that the fused AOD dataset gained a similar accuracy, with R^2 of 0.77 and RMSE of 0.25 (Fig. S9), to the combined imputations shown in Fig. 5 (a).

4. Discussion

4.1. Strengths of the proposed imputation modeling

It is difficult to impute spatiotemporal high-resolution AOD towards full coverage, because the target variable (i.e., MAIAC retrieved AOD) highly varies in space over time. The relationship between the target AOD and its covariates is thus too complicated to characterize using a simple model (e.g., linear regression). Such an issue becomes more complex and intractable when filling gaps for an extensive area with heterogeneous surface conditions during a very long period, where the

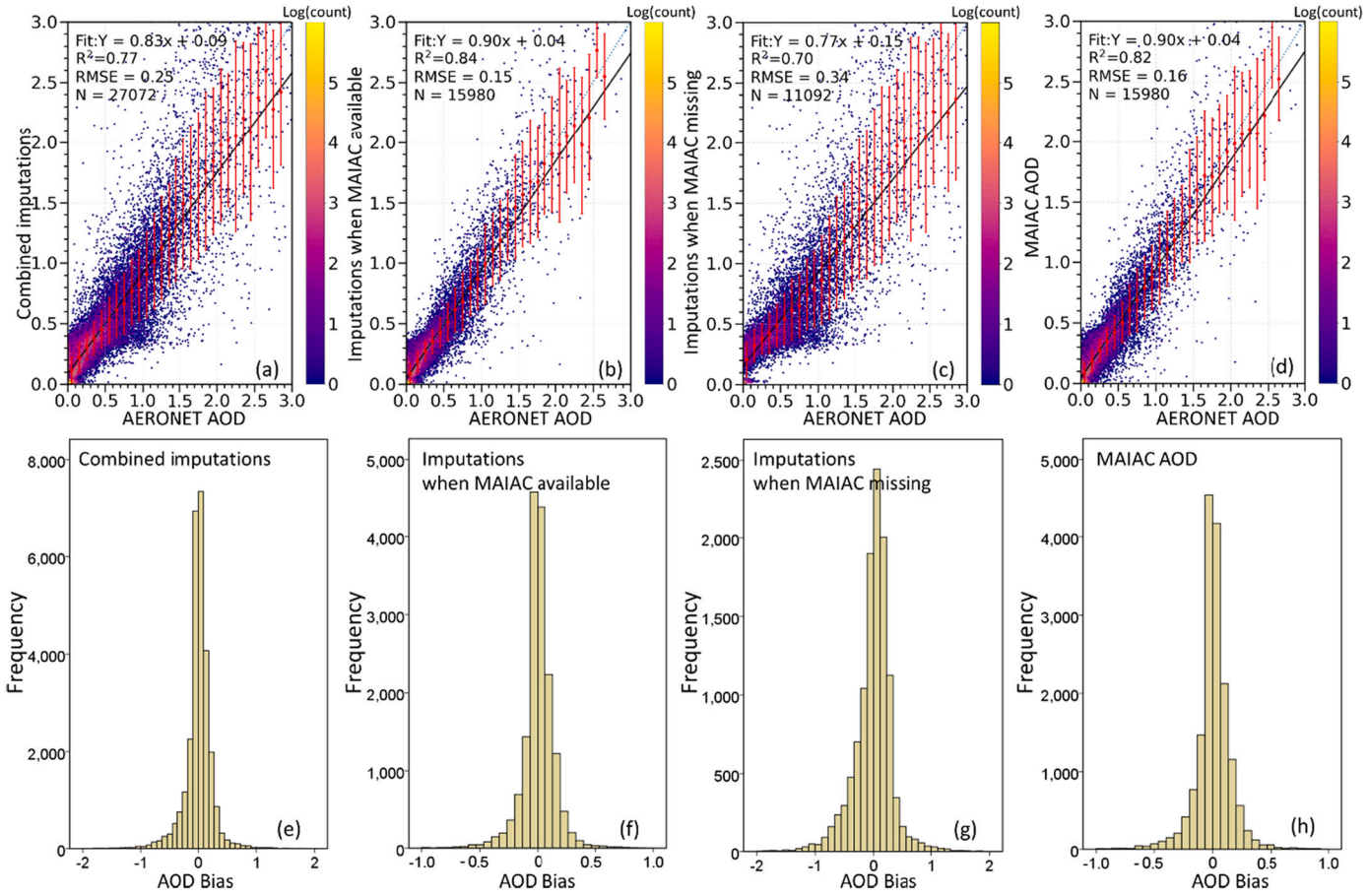

Fig. 5. (a-d) Scatterplots and (e-h) histograms of imputed vs. AERONET AODs. The red dots with error bars in the upper panel represent the mean imputed AOD separated by 0.1 AERONET bins and the corresponding one standard deviation; the black line represents the linear regression line between the imputed and AERONET AODs, and the blue dotted line is the one-to-one line.

spatiotemporal variability increases dramatically. Most previous studies involved in filling AOD gaps developed imputation models either in small study areas or over short periods, where less spatiotemporal variability was handled. However, despite reduced variability, the model performance remains to be improved, with validated R^2 values of 0.18–0.52 against AERONET observations (Goldberg et al., 2019; Lv et al., 2017; Xiao et al., 2017; Zhao et al., 2019) or incomplete imputation coverage (Jinnagara Puttaswamy et al., 2014; Wang et al., 2019; Yang and Hu, 2018; Zhao et al., 2021).

Here, we proposed a random forest-based imputation modeling framework, coupling an adaptive process of optimizing hyperparameters and a time-stratified sampling method. Using a set of spatiotemporally varying predictors (EAC4 AOD, meteorology, surface conditions, MAIAC monthly AOD, spatial coordinates, and temporal dummy index), our daily imputation modeling strategy efficiently resolved the spatiotemporal variability within the data and entirely made up AOD gaps for each modeling day. Compared with satellite and ground-level observations, our modeling method arguably achieved state-of-the-art performance, improving daily aerosol data availability to 100% while maintaining high accuracy of AOD imputations (hold-out testing R of 0.988 against the target MAIAC retrieved AOD and validation R^2 of 0.77 against the ground-truth AERONET AOD). Then, we applied the developed models to generate daily AOD imputations across China from 2003 to 2020 and finally obtained the fused daily 1-km AOD dataset by combining them with MAIAC original retrievals.

A hybrid dataset, including satellite observations and simulations, has been employed to expand the spatiotemporal coverage of MAIAC retrieved AOD data. The simulated AOD (EAC4 AOD), which has coarse spatial resolution but fine temporal interval, was correlated well with MAIAC AOD with an overall correlation coefficient r of 0.749. EAC4 can figure out the overview of the daily aerosol pattern and was used as a primary predictor in imputation modeling. Temporally varying predictors such as meteorological variables helped each daily model resolve the local temporal variations in the data. Surface condition variables such as topography, NDVI, and coordinates have been proven in good agreement with aerosol distribution (He et al., 2019; Xiao et al., 2017), which were used as spatially varying predictors to help each model capture local variations of aerosol loading. Monthly MAIAC maps were incorporated as model inputs for areas with enough valid AOD values in a natural month, helping the model represent more spatial gradients of aerosol loading. Cloud fraction was also included in model development to represent the interaction between aerosol and cloud, assisting the spatial inference (Fig. 6). Fig. 6 exemplifies our fused AOD vs. MAIAC original retrievals on 26 Jan 2013, a cloudy day with high particulate pollution observed by surface monitors. Comparison results of our imputations with MAIAC retrievals and AERONET measurements indicate that the imputed AOD values are reliable extensions on cloudy days when a large amount of AOD retrievals are missing.

An adaptive modeling framework was developed to optimize the hyperparameters and train the samples, significantly saving computational costs. Compared with those models that only included samples from the modeling day, our daily imputation models were trained on temporally stratifying samples, which can capitalize on the temporal dependence within a short-term period. With these advances, our proposed framework has successfully bridged gaps left by the MAIAC aerosol retrieval algorithm over China for a long period (2003−2020), at

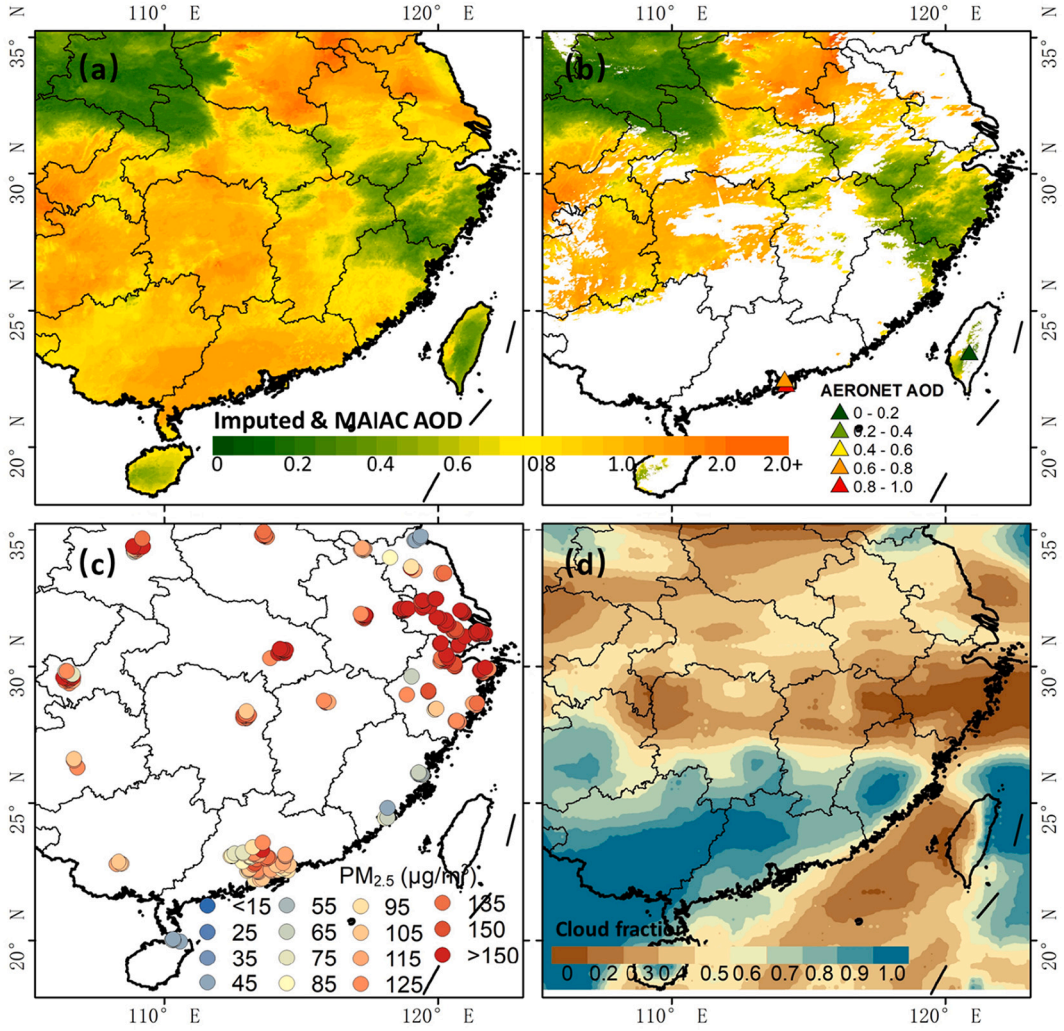

Fig. 6. Fused AOD on a cloudy day with high $PM_{2.5}$ pollution observed by ground-level air quality monitors (26 Jan 2013): (a) spatial distribution of fused AOD values, (b) MAIAC AOD retrievals, (c) daily mean $PM_{2.5}$ concentrations by national air quality monitoring network, and (d) total cloud fraction by the ECMWF reanalysis.

Table 2
Comparison of validation results in this study with previously reported results in the literature.

Validation results in previous studies								Validation results in this study		
Reference	Spatiotemporal parameters		Imputation method	Validation						
	Coverage	Scale		Method	r	R^2	RMSE	r	R^2	RMSE
Bai et al., 2022	China 2000–2020	Daily 1 km	Tensor-flow-based data fusion	With AERONET	0.91	–	0.21	0.88	0.77	0.25
Li, 2021	China 2015–2018	Daily 1 km	Deep learning	With AERONET	–	0.78	0.27	–	0.84	0.23
Li, 2020	BTH 2015	Daily 1 km	Deep learning	With AERONET	–	0.82	0.21	–	0.78	0.33
Li et al., 2020	California, US 2000–2016	Weekly 1 km	Deep learning	With AERONET	–	0.45	0.06	–	0.77	0.25
Goldberg et al., 2019	Eastern US 2008	Daily 1 km	Regression + IDW*	With AERONET	–	0.52	–	–	0.79	–
Jiang et al., 2021	China 2018.3–2019.2	Hourly 1 km	Random forest	With AERONET	0.65		0.20	0.81		0.21
Xiao et al., 2017	YRD, China 2014	Daily 1 km	GAM*	With AERONET	–	0.44	–	–	0.69	–
Zhao et al., 2019	BTH, China 2010–2016	Daily 3 km	Random forest	With AERONET	–	0.36	–	–	0.80	–
Lv et al., 2017	BTH, China 2014	4 km Daily	Regression + Kriging	5-fold cross validation	–	0.39	–	–	0.98	–

* GAM - Generalized additive model; IDW – Inverse distance weighting interpolation.

a 1-km grid and a daily timescale.

4.2. Comparison with other imputation models

A few studies have developed imputation methods to improve the spatial coverage of satellite high-resolution AOD to approximately 100% at the daily timescale. However, most previous studies did not report validation results of imputations because the imputed AOD was an intermediate output and used as an explanatory variable for the subsequent $PM_{2.5}$ modeling. Table 2 compared validation results from our fused 1-km AOD dataset and those reported by previous studies. To make the comparison comparable, we extracted subsets from our validation results to spatiotemporally match the study domain of previous research (Goldberg et al., 2019; Jiang et al., 2021; Li, 2021; Zhao et al., 2019). In general, validation results in Table 2 demonstrate that our gap-filled approach performed better or at least comparable to those in the literature.

Undoubtedly, our proposed imputation approach outperforms those relying on the generalized additive model (Xiao et al., 2017) and interpolation methods (Goldberg et al., 2019; Lv et al., 2017), significantly increasing R^2 values by 0.25–0.59 for the validation against AERONET measurements. Li (2021) applied a deep learning approach to impute 1-km AOD over China from 2015 to 2018, achieving an overall R^2 and RMSE values of 0.78 and 0.27 in comparison with ground-truth values; with a matched spatiotemporal coverage, validation R^2 and RMSE values of this study were better (0.84 and 0.23). However, it is worth noting that the collocation strategy may influence validation results against AERONET AOD (e.g., applying different spatial windows and temporal intervals to match AERONET and imputed AODs), despite not being significant (Zhang et al., 2019). Li et al. (2020) also used a deep learning method to improve spatiotemporal coverage of MAIAC weekly AOD values and slightly enhanced the accuracy (R^2 = 0.45, RMSE = 0.06) compared to MAIAC original retrievals (R^2 = 0.44, RMSE = 0.06). The significantly lower RMSE value was primarily due to lower aerosol loading over the modeling area (overall mean AOD = 0.084 vs. 0.230 in this study). Bai et al. (2022) developed a TensorFlow-based data fusion method to fill the gaps left by MAIAC AOD retrievals and achieved a little higher validation accuracy than this study; however, in addition to those publicly available external data, they also employed some input data (e.g., visibility data from high-density weather stations) that are not readily available for the public.

4.3. Imputed AODs vs. MAIAC retrievals at longer timescales

Since the proportion of AOD missing values vary spatiotemporally, this section compared the uncertainty and representativeness between MAIAC AOD retrievals and our fused full-coverage values at longer (monthly/seasonal/annual) timescales. MAIAC month-aggregated AODs were calculated by averaging all available daily retrievals; due to the missingness issue, the monthly mean AOD may not cover the entire study area. Fig. 7 shows that our monthly fused AOD values have lower uncertainty than MAIAC monthly aggregated value, with better R^2 and RMSE values against AERONET measurements (R^2 = 0.87, RMSE = 0.13 vs. R^2 = 0.80, RMSE = 0.16). Since both AERONET (due to cloudy skies) and MAIAC AOD cannot have 100% availability each month, we also examined the impact of various missing rates on the statistical errors. Fig. S10 and Fig. S11 demonstrate that when the proportions of monthly AERONET and MAIAC available data exceeded 70%, the validation R^2 for MAIAC (0.88) was close to the fused one (0.87), but the sample size due to the MAIAC missingness reduced significantly to 99 from 393. Thus, our fused monthly AODs with full coverage have superior precision than MAIAC monthly values with missing values.

Seasonal mean values were also computed and compared, as shown in Table 3. AERONET measurements and fused values of AOD had 100% availability, and the missing rates for MAIAC AOD retrievals were present. Overall, due to the missingness issue, the time-aggregated MAIAC AOD retrievals underestimated aerosol loading in the atmosphere, with 26% of the mean value lower than AERONET's; by contrast, the deviation for our full-coverage AOD imputations was only 1%. The

Table 3
The seasonal mean AOD values of AERONET measurements and imputations, as well as the missing rates of MAIAC AOD retrievals by seasons in the sample dataset from 2003 to 2020.

Season	No. of measurements	Mean of measurements	Mean of fused AOD	Mean of MAIAC AOD	Missing rate of MAIAC (%)*
Spring	7844	0.516	0.499	0.384	41
Summer	6104	0.573	0.620	0.401	54
Autumn	6578	0.426	0.432	0.345	37
Winter	6546	0.369	0.360	0.280	32
Total	27,072	0.471	0.476	0.348	41

* Missing rate of MAIAC = 1-No. of AERONET measurements matched with valid MAIAC AOD/No. of all AERONET measurements.

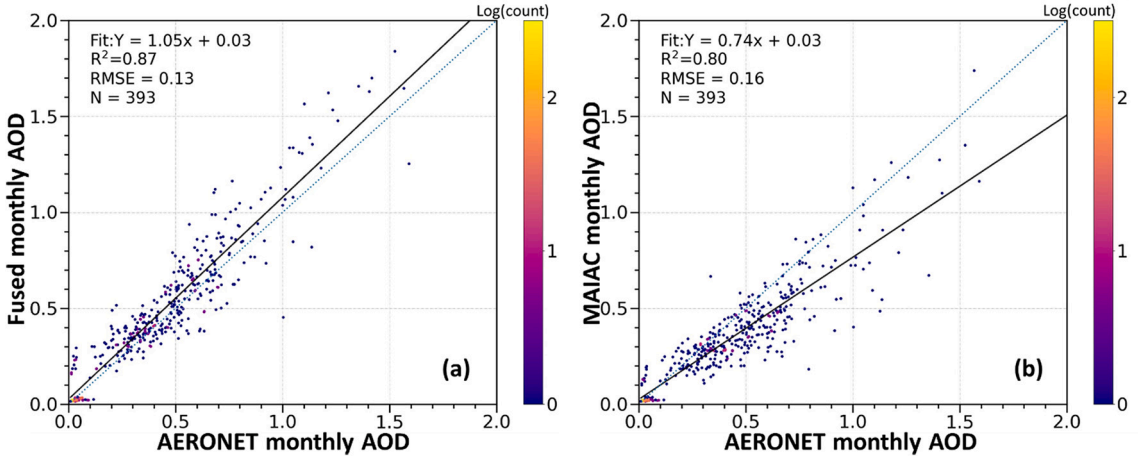

Fig. 7. Comparison results for (a) AERONET and fused monthly mean AOD by the present work and (b) AERONET-MAIAC monthly AODs during 2003–2020. Because AERONET aerosol measurements had missing values, the monthly mean AERONET AOD required at least 70% of available daily measurements each month to guarantee its representativeness. Due to the missingness of MAIAC AOD retrievals, the monthly mean values of MAIAC AOD retrievals show larger uncertainty than the monthly fused values.

underestimation in summer was up to 30% for the seasonally aggregated MAIAC AOD retrievals, while the maximum missing rate was observed (54%). Consequently, the absence of satellite retrieved AOD leads to an underestimation of the time-aggregated aerosol loads in the total atmosphere, which would bias the understanding of the long-term characteristics of atmospheric aerosol particles and necessitate improving the spatiotemporal coverage of AOD data.

4.4. National AOD maps and evolution pattern

Using satellite aerosol retrievals, previous studies have investigated the global and regional variations of AOD and reported long-term trends in China: aerosol loading went up in the first decade of 21 century but declined in the long run (Filonchyk et al., 2019; Gao et al., 2018; Guo et al., 2011; He et al., 2019; Mao et al., 2014; Sogacheva et al., 2018). Such long-term trends observed by previous studies are overall comparable to those by the present work (Fig. S12-S14), considering the differences in study regions and timespans. However, monthly and seasonal comparison results in Fig. 7 and Table 3 clearly show that if the missingness of AOD was left unsolved, spatial patterns calculated only by MAIAC AOD retrievals significantly underestimate the columnar AOD in China on seasonal and yearly timescales. Fig. 8 profiles the spatial pattern of differences between our fused AOD and MAIAC retrievals over the past 18 years. It can be clearly seen that the underestimation based on MAIAC retrievals on longer timescales primarily occurred in eastern China, particularly in the NCP, western Hubei-Hunan, and SCB regions, with differences as large as 0.2 and even larger.

To investigate the evolution and trend of aerosol loading, we calculated annual mean AOD anomalies and applied them in the spatiotemporal analysis. The annual anomaly map was calculated by subtracting the 18-year average from the corresponding annual average for each grid cell. Fig. 9 presents the 18-year mean AOD map and annual anomaly maps from 2003 to 2020 based on our fused AOD data. The overall AOD throughout the study domain is 0.34, with an std. of 0.24. In general, the multi-year mean AOD map profiles the spatial variation of AOD across China, with high values (>0.60) clustered in eastern China and gradually declining to the west of the country. The AOD values were particularly high in the NCP, SCB, and Hubei-Hunan regions, with most areas having 18-year mean values >0.80. In the past eighteen years, atmospheric aerosol loads in western China were relatively lower, with values of <0.35, except for the Taklimakan Desert (multi-year mean values of 0.3–0.55), where dust storms frequently occurred in spring (Liu et al., 2019a).

The annual anomaly maps, around the multi-year mean AOD in Fig. 9, demonstrate the temporal evolution of aerosol loading over China from 2003 to 2020. The evolution pattern shows the significant regional difference. In 2003, strong positive anomalies (>0.05) observed in northeastern China indicate that the aerosol loading during this year was higher compared to the 18-year mean values; during 2004–2005, the AOD anomalies in the northeast and the NCP region converted to negative, between −0.12 and 0, suggesting the annual mean AOD decreased during the two years. The prevalent positive anomalies in 2003 in northeastern China most likely resulted from several big forest fires in eastern Russia (https://earthobservatory.nasa.gov/images/related/11785/forest-fires-in-eastern-russia). In the three years of 2003–2005, the extent of positive anomalies became larger over southern China, from Jiangxi province to entire South China. Between 2006 and 2012, strong and continuous positive anomalies (>0.05) were found over most of eastern China. The increase in aerosol loading in these years mainly resulted from urbanization and industrialization associated with relatively loose emission policies (Ma et al., 2019; Sogacheva et al., 2018). Towards 2013, positive anomalies changed to weak negative over southeastern China and became less positive (<0.05) in other southern areas, while remaining high (>0.05) in most of the NCP region; however, in 2014, significant positive anomalies prevailed in eastern China. Starting from 2015, strong negative AOD anomalies (<−0.05) were observed in SCB, followed by weak negative deviations in other eastern China in 2016. After 2017, stronger negative anomalies (<−0.10) gradually enlarged over extensive areas of eastern China, and anomalies were significantly negative (<−0.05) over the northeast of the country. The sudden drop in AOD in the recent six years was primarily related to the ongoing stringent-ever emission control and

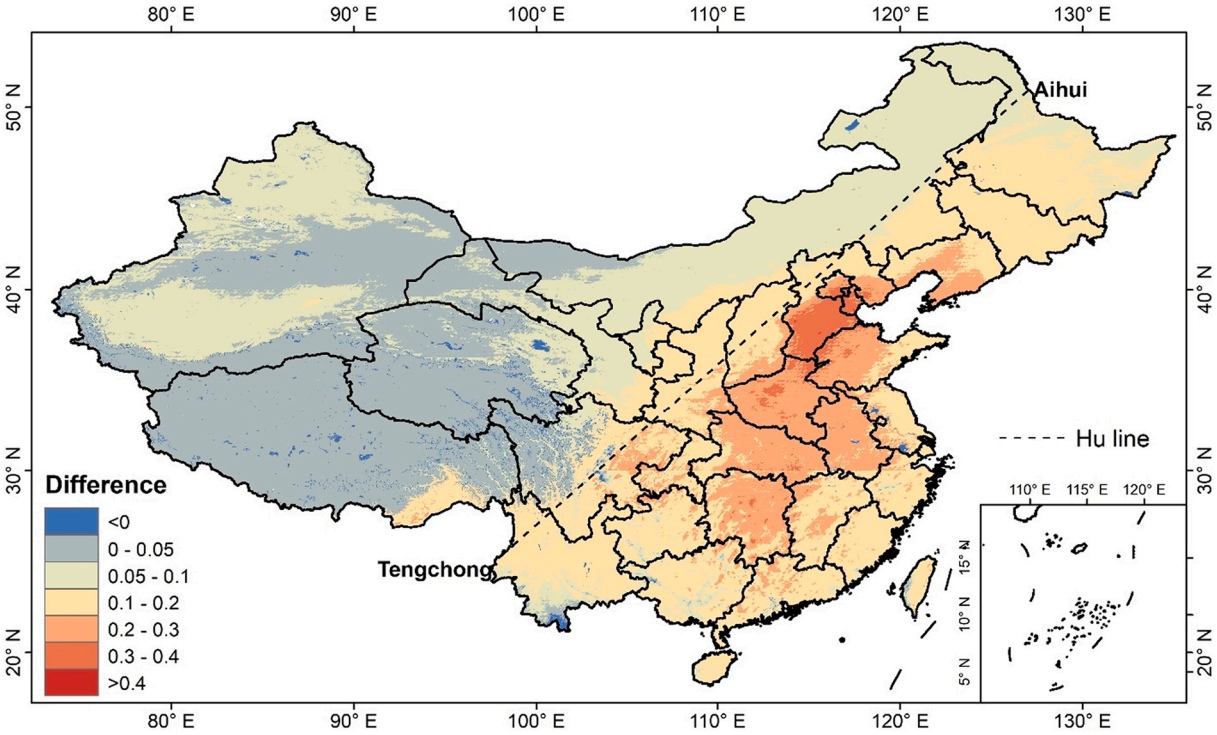

Fig. 8. Spatial distribution of differences between multi-year mean fused and MAIAC AOD in 2003–2020.

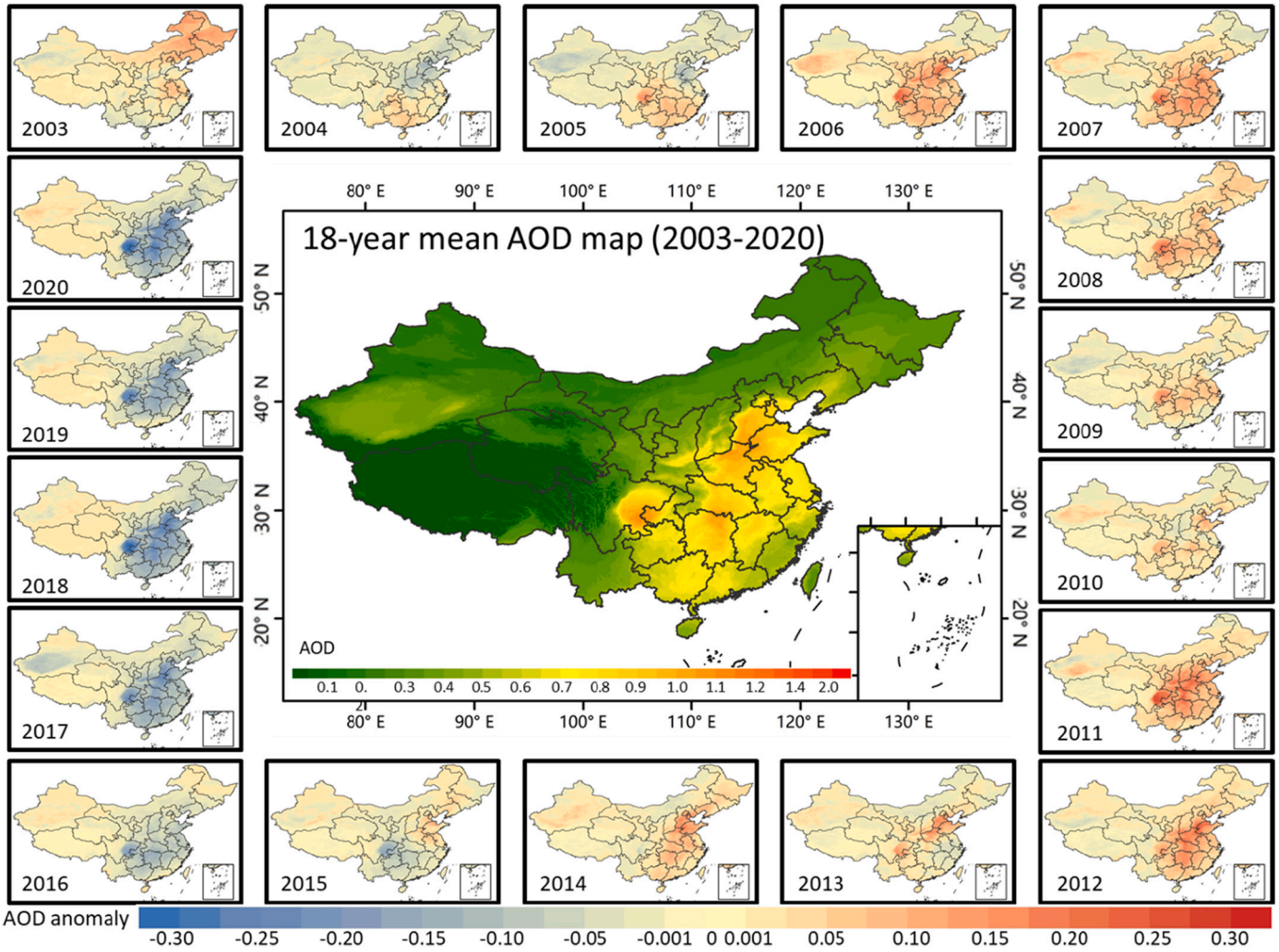

Fig. 9. Multi-year mean 1-km AOD map (middle) and annual AOD anomaly maps (surrounding) for 2003 to 2020 based on the fused full-coverage AOD data.

air pollution prevention regulations implemented since 2013 (Ma et al., 2019; Sogacheva et al., 2018). Unlike the variable AOD anomaly evolution pattern in eastern China, AOD behaviors in western China were different: the 18-year mean AOD map seems fairly represent levels of aerosol loading throughout the study timespan, only with small deviations between −0.025 and 0.025 around the averages. The Taklimakan Desert was an exception, where AOD anomalies may significantly fluctuate from zero to a large value of ±0.05.

4.5. Applications of imputed AOD data

Exploring short-term variations of atmospheric aerosol loading requires accurate aerosol data on fine (e.g., daily) timescales. For example, the long-lasting events of extreme particle pollution frequently occur in eastern China, especially in the NCP region in winter, which has drawn widespread attention in the scientific community (Wang et al., 2014). However, it is beyond the ability of MAIAC AOD retrievals with incomplete spatiotemporal coverage (Fig. 10 c-d) because the spatiotemporally discontinuous data cannot present the evolution of severe haze episodes for identifying the forming mechanism. Our fused aerosol data that offer time series of spatially continuous pictures of aerosol particles countrywide thus can provide an opportunity to analyze the process and evolution of such events.

Other important applications of atmospheric aerosol data include the quantitative analysis of impact factors on aerosol particles (He et al., 2019; Wang et al., 2017) and the estimation of ambient concentrations of $PM_{2.5}$ (or PM_{10}) for environmental management and health-related studies (He et al., 2021b; Ma et al., 2019; Meng et al., 2021; Wei et al., 2021). However, before this study, such countrywide applications could only use the spatiotemporally discontinuous AOD, which poses a similar dilemma: the unavailability (He et al., 2021b; Wei et al., 2021) or additional uncertainty (Meng et al., 2021) induced by the missingness of AOD. It is beyond the scope of the present work to quantify the attributable factors on the AOD variations or estimate ground-level particulate concentrations; however, supported by the reliable aerosol data in this study, future research on both country and regional scope in this field can be warranted.

4.6. Limitations

The major limitation of this work lies in the uncertainty in our imputations. The data used as the dependent variable in modeling was derived from MAIAC 1-km aerosol retrievals that themselves have uncertainty due to limitations of the MAIAC retrieval algorithm (Lyapustin et al., 2018), even though it is the best-available aerosol data so far. The present hold-out validation by randomly separating the sample dataset into training and testing folders may enhance the model performance because there is a correlation between the training and testing samples. However, considering the massive computational burden, we did not take the spatial cross-validation method to quantify the model errors in our imputations. The comparisons with AERONET ground-level measurements, which were not used in model development, assisted in

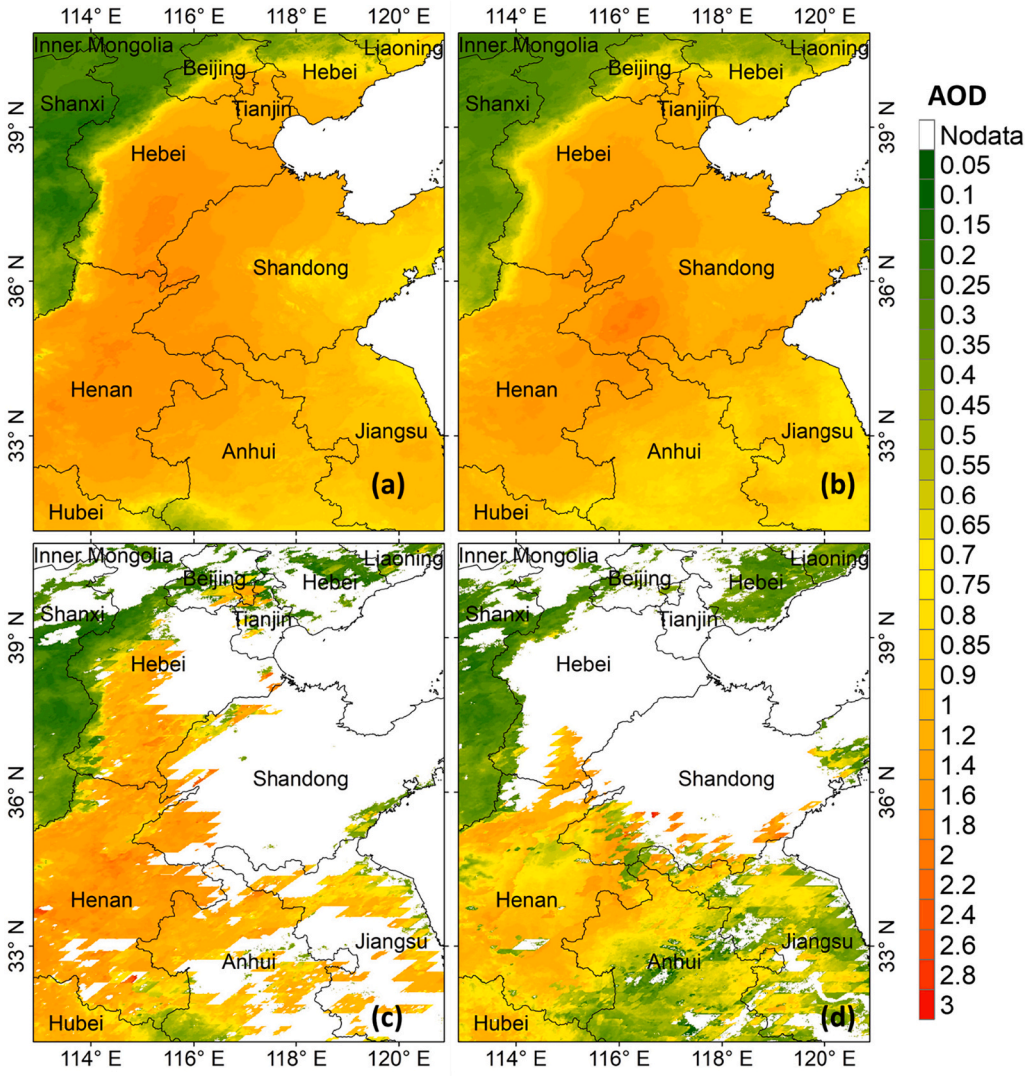

Fig. 10. Two typical long-lasting episodes of extremely severe haze occurred in central and eastern China. Spatial maps of the (a, b) fused and (c, d) MAIAC AOD values averaged during in 2013: (a, c) are for 9–15 January and (b, d) for 25–31 January. Due to missing values, the MAIAC data fail to represent the complete pictures for the two haze events and cannot be used to study the forming mechanism of haze, but it is possible for the daily fused data generated by this study.

quantitatively assessing the uncertainty in the results during 2003–2019, but this may not be very representative in space because only a limited number of monitoring stations are sparsely located in the study area.

5. Conclusions

This study developed a spatiotemporal high-resolution imputation modeling approach based on random forest models to elucidate robust daily relationships between MAIAC retrieved AOD and its covariates for China and generate a daily 1-km AOD dataset with full coverage from 2003 to 2020. Cloud-aerosol interaction, meteorological parameters, surface conditions, and spatiotemporal dummy indicators were incorporated to help the model resolve the spatiotemporal variability in the data and improve the MAIAC retrieved AOD coverage to 100% of completeness. Validation and comparison results demonstrate that the imputation models (overall hold-out validation correlation coefficient/RMSE of 0.988/0.036 and validation R^2/RMSE against surface AOD measurements of 0.77/0.25) achieved better accuracy than most imputation strategies adopted in previous studies. Comparison results among full-coverage AOD imputations, MAIAC original retrievals, and AERONET aerosol measurements indicate that filling satellite AOD gaps can improve the accuracy of time-aggregated satellite AOD values. Due to the missingness of the satellite AOD, the monthly/seasonal/yearly aerosol loading in most of China tends to be underestimated, particularly in summer and over the NCP region (the amount of underestimation >0.2). The long-term full-coverage data of spatially ubiquitous AOD can support the scientific research on climate change, air quality, and epidemiology associated with atmospheric aerosols in China, on both national and local scales. The final AOD dataset is available at https://dataverse.harvard.edu/dataverse/atmospheric_data_by_WHUT.

CRediT authorship contribution statement

Qingqing He: Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Weihang Wang:** Data curation, Software, Visualization, Writing – review & editing. **Yimeng Song:** Writing – review & editing. **Ming Zhang:** Writing – review & editing. **Bo Huang:** 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.

Data availability

The final AOD dataset is publicly available at https://dataverse.harvard.edu/dataverse/atmospheric_data_by_WHUT.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Grant NO. 41901324, 42201369) and the Fundamental Research Funds for the Central Universities, China (WUT: 223108007).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosres.2022.106481.

References

Abuelgasim, A., Bilal, M., Alfaki, I.A., 2021. Spatiotemporal variations and long term trends analysis of aerosol optical depth over the United Arab Emirates. Remote Sens. Appl. Soc. Environ. 23, 100532.
Bai, K., Li, K., Ma, M., Han, D., 2022. LGHAP: the long-term Gap-free High-resolution Air Pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion. Earth Syst. Sci. Data 14, 907–927.
Bi, J., Belle, J.H., Wang, Y., Lyapustin, A.I., Wildani, A., Liu, Y., 2019. Impacts of snow and cloud covers on satellite-derived PM2. 5 levels. Remote Sens. Environ. 221, 665–674.
Brauer, M., Freedman, G., Frostad, J., Van Donkelaar, A., Martin, R.V., Dentener, F., Dingenen, R.V., Estep, K., Amini, H., Apte, J.S., 2016. Ambient air pollution exposure estimation for the global burden of disease 2013. Environ. Sci. Technol. 50, 79–88.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., Van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934.
Chen, Z.-Y., Jin, J.-Q., Zhang, R., Zhang, T.-H., Chen, J.-J., Yang, J., Ou, C.-Q., Guo, Y., 2020. Comparison of Different Missing-Imputation Methods for MAIAC (Multiangle Implementation of Atmospheric Correction) AOD in estimating Daily PM2. 5 Levels. Remote Sens. 12, 3008.
Chudnovsky, A., Tang, C., Lyapustin, A., Wang, Y., Schwartz, J., Koutrakis, P., 2013a. A critical assessment of high-resolution aerosol optical depth retrievals for fine particulate matter predictions. Atmos. Chem. Phys. 13, 10907–10917.
Chudnovsky, A.A., Kostinski, A., Lyapustin, A., Koutrakis, P., 2013b. Spatial scales of pollution from variable resolution satellite imaging. Environ. Pollut. 172, 131–138.
Collaborators, GBD, 2018. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet (London, England) 392, 1923.
Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. Environ. Sci. Technol. 50, 4712–4721.
Filonchyk, M., Yan, H., Zhang, Z., Yang, S., Li, W., Li, Y., 2019. Combined use of satellite and surface observations to study aerosol optical depth in different regions of China. Sci. Rep. 9, 1–15.
Gao, L., Chen, L., Li, J., Heidinger, A.K., Xu, X., Qin, S., 2018. A long-term historical aerosol optical depth data record (1982–2011) over China from AVHRR. IEEE Trans. Geosci. Remote Sens. 57, 2467–2480.
Goldberg, D.L., Gupta, P., Wang, K., Jena, C., Zhang, Y., Lu, Z., Streets, D.G., 2019. Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM2. 5 concentrations at 1 km resolution in the Eastern United States. Atmos. Environ. 199, 443–452.
Guo, J., Zhang, X., Wu, Y., Zhaxi, Y., Che, H., La, B., Wang, W., Li, X., 2011. Spatio-temporal variation trends of satellite-based aerosol optical depth in China during 1980–2008. Atmos. Environ. 45, 6802–6811.
He, Q., Gu, Y., Zhang, M., 2019. Spatiotemporal patterns of aerosol optical depth throughout China from 2003 to 2016. Sci. Total Environ. 653, 23–35.
He, Q., Gao, K., Zhang, L., Song, Y., Zhang, M., 2021a. Satellite-derived 1-km estimates and long-term trends of PM2. 5 concentrations in China from 2000 to 2018. Environ. Int. 156, 106726.
He, Q., Zhang, M., Song, Y., Huang, B., 2021b. Spatiotemporal assessment of PM2. 5 concentrations and exposure in China from 2013 to 2017 using satellite-derived data. J. Clean. Prod. 286, 124965.
Holben, B., Eck, T., Slutsker, I., Tanré, D., Buis, J., Setzer, A., Vermote, E., Reagan, J., Kaufman, Y., Nakajima, T., 1998. AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization. Remote Sens. Environ. 66, 1–16.
Huang, C., Hu, J., Xue, T., Xu, H., Wang, M., 2021. High-Resolution Spatiotemporal Modeling for Ambient PM2. 5 Exposure Assessment in China from 2013 to 2019. Environ. Sci. Technol. 55, 2152–2162.
Inness, A., Ades, M., Agusti-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J.J., Engelen, R., Eskes, H., Flemming, J., 2019. The CAMS reanalysis of atmospheric composition. Atmos. Chem. Phys. 19, 3515–3556.
Jiang, T., Chen, B., Nie, Z., Ren, Z., Xu, B., Tang, S., 2021. Estimation of hourly full-coverage PM2. 5 concentrations at 1-km resolution in China using a two-stage random forest model. Atmos. Res. 248, 105146.
Jinnagara Puttaswamy, S., Nguyen, H.M., Braverman, A., Hu, X., Liu, Y., 2014. Statistical data fusion of multi-sensor AOD over the continental United States. Geocarto Int. 29, 48–64.
Kaufman, Y.J., Tanré, D., Boucher, O., 2002. A satellite view of aerosols in the climate system. Nature 419, 215–223.
Kianian, B., Liu, Y., Chang, H.H., 2021. Imputing Satellite-Derived Aerosol Optical Depth using a Multi-Resolution Spatial Model and Random Forest for PM2.5 Prediction. Remote Sens. 13.
Kloog, I., Nordio, F., Coull, B.A., Schwartz, J., 2012. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2. 5 exposures in the Mid-Atlantic states. Environ. Sci. Technol. 46, 11913–11921.
Levy, R., Mattoo, S., Munchak, L., Remer, L., Sayer, A., Hsu, N., 2013. The Collection 6 MODIS aerosol products over land and ocean. Atmos. Meas. Tech. Discuss 6, 159–259.
Li, L., 2020. A robust deep learning approach for spatiotemporal estimation of satellite AOD and PM2. 5. Remote Sens. 12, 264.
Li, L., 2021. High-Resolution Mapping of Aerosol Optical Depth and Ground Aerosol Coefficients for mainland China. Remote Sens. 13, 2324.
Li, L., Franklin, M., Girguis, M., Lurmann, F., Wu, J., Pavlovic, N., Breton, C., Gilliland, F., Habre, R., 2020. Spatiotemporal imputation of MAIAC AOD using deep learning with downscaling. Remote Sens. Environ. 237, 111584.
Li, S., Chen, L., Tao, J., Han, D., Wang, Z., Su, L., Fan, M., Yu, C., 2012. Retrieval of aerosol optical depth over bright targets in the urban areas of North China during winter. Sci. China Earth Sci. 55, 1545–1553.
Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017. Estimating ground-level PM2. 5 by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44, 11,985-911,993.
Li, Z., Xia, X., Cribb, M., Mi, W., Holben, B., Wang, P., Chen, H., Tsay, S.C., Eck, T., Zhao, F., 2007. Aerosol optical properties and their radiative effects in northern China. J. Geophys. Res.-Atmos. 1984–2012, 112.
Lin, C., Li, Y., Lau, A.K., Deng, X., Tim, K., Fung, J.C., Li, C., Li, Z., Lu, X., Zhang, X., 2016. Estimation of long-term population exposure to PM 2.5 for dense urban areas using 1-km MODIS data. Remote Sens. Environ. 179, 13–22.
Liu, J., Wang, R., Zhao, Y., Yang, Y., 2019a. A 40,000-year record of aridity and dust activity at Lop Nur, Tarim Basin, northwestern China. Quat. Sci. Rev. 211, 208–221.
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.
Lops, Y., Pouyaei, A., Choi, Y., Jung, J., Salman, A.K., Sayeed, A., 2021. Application of a partial convolutional neural network for estimating geostationary aerosol optical depth data. Geophys. Res. Lett. 48, e2021GL093096.
Lv, B., Hu, Y., Chang, H.H., Russell, A.G., Cai, J., Xu, B., Bai, Y., 2017. Daily estimation of ground-level PM2. 5 concentrations at 4 km resolution over Beijing-Tianjin-Hebei by fusing MODIS AOD and ground observations. Sci. Total Environ. 580, 235–244.
Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., Reid, J.S., 2011. Multi-Angle Implementation of Atmospheric Correction (MAIAC): 2. Aerosol Algorithm. J. Geophys. Res. 116, D03211.
Lyapustin, A., Wang, Y., Korkin, S., Huang, D., 2018. MODIS Collection 6 MAIAC algorithm. Atmos. Meas. Tech. 11.
Ma, X., Wang, J., Yu, F., Jia, H., Hu, Y., 2016. Can MODIS AOD be employed to derive PM2. 5 in Beijing-Tianjin-Hebei over China? Atmos. Res. 181, 250–256.
Ma, Z., Liu, R., Liu, Y., Bi, J., 2019. Effects of air pollution control policies on PM 2.5 pollution improvement in China from 2005 to 2017: A satellite-based perspective. Atmos. Chem. Phys. 19, 6861–6877.
Mao, K., Ma, Y., Xia, L., Chen, W.Y., Shen, X., He, T., Xu, T., 2014. Global aerosol change in the last decade: an analysis based on MODIS data. Atmos. Environ. 94, 680–686.
Mei, Linlu, Strandgren, Johan, Rozanov, Vladimir, Vountas, Marco, Burrows, John P., Wang, Yujie, 2019. A study of the impact of spatial resolution on the estimation of particle matter concentration from the aerosol optical depth retrieved from satellite observations. Int. J. Remote Sens. 40 (18), 7084–7112.
Meng, X., Liu, C., Zhang, L., Wang, W., Stowell, J., Kan, H., Liu, Y., 2021. Estimating PM2. 5 concentrations in Northeastern China with full spatiotemporal coverage, 2005-2016. Remote Sens. Environ. 253, 112203.
Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: Lines that connect. J. Air Waste Manage. Assoc. 56, 709–742.
Pu, Q., Yoo, E.-H., 2021. Ground PM2. 5 prediction using imputed MAIAC AOD with uncertainty quantification. Environ. Pollut. 274, 116574.
Remer, L., Mattoo, S., Levy, R., Munchak, L., 2013. MODIS 3 km aerosol product: algorithm and global perspective. Atmos. Meas. Tech. 6, 1829–1844.
Schneider, R., Vicedo-Cabrera, A.M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., Kloog, I., Reis, S., Vieno, M., Gasparrini, A., 2020. A satellite-based spatio-temporal machine learning model to reconstruct daily PM2. 5 concentrations across Great Britain. Remote Sens. 12, 3803.
Sogacheva, L., Rodriguez, E., Kolmonen, P., Virtanen, T.H., Saponaro, G., Leeuw, G.d., Georgoulias, A.K., Alexandri, G., & Kourtidis, K., 2018. Spatial and seasonal

variations of aerosols over China from two decades of multi-satellite observations–part 2: AOD time series for 1995–2017 combined from ATSR ADV and MODIS C6. 1 and AOD tendency estimations. Atmos. Chem. Phys. 18, 16631–16652.
Song, W., Jia, H., Huang, J., Zhang, Y., 2014. A satellite-based geographically weighted regression model for regional PM 2.5 estimation over the Pearl River Delta region in China. Remote Sens. Environ. 154, 1–7.
Song, Z., Fu, D., Zhang, X., Han, X., Song, J., Zhang, J., Wang, J., Xia, X., 2019. MODIS AOD sampling rate and its effect on PM2. 5 estimation in North China. Atmos. Environ. 209, 14–22.
Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., De Hoogh, K., De'Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., 2019. Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179.
Tao, M., Chen, L., Wang, Z., Tao, J., Che, H., Wang, X., Wang, Y., 2015. Comparison and evaluation of the MODIS Collection 6 aerosol data in China. J. Geophys. Res.-Atmos. 120, 6992–7005.
Van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., Lyapustin, A., Sayer, A.M., Winker, D.M., 2016. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. Environ. Sci. Technol. 50, 3762–3772.
Wang, C., Wang, C., Myint, S.W., Wang, Z.-H., 2017. Landscape determinants of spatio-temporal patterns of aerosol optical depth in the two most polluted metropolitans in the United States. Sci. Total Environ. 609, 1556–1565.
Wang, Y., Yao, L., Wang, L., Liu, Z., Ji, D., Tang, G., Zhang, J., Sun, Y., Hu, B., Xin, J., 2014. Mechanism for the formation of the January 2013 heavy haze pollution episode over central and eastern China. Sci. China Earth Sci. 57, 14–25.
Wang, Y., Yuan, Q., Li, T., Shen, H., Zheng, L., Zhang, L., 2019. Large-scale MODIS AOD products recovery: Spatial-temporal hybrid fusion considering aerosol variation mitigation. ISPRS J. Photogramm. Remote Sens. 157, 1–12.
Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., Cribb, M., 2021. Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Remote Sens. Environ. 252, 112136.
Xiao, Q., Wang, Y., Chang, H.H., Meng, X., Geng, G., Lyapustin, A., Liu, Y., 2017. Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sens. Environ. 199, 437–446.
Xiao, Q., Geng, G., Cheng, J., Liang, F., Li, R., Meng, X., Xue, T., Huang, X., Kan, H., Zhang, Q., 2021. Evaluation of gap-filling approaches in satellite-based daily PM2. 5 prediction models. Atmos. Environ. 244, 117921.
Xu, H., Guang, J., Xue, Y., De Leeuw, G., Che, Y., Guo, J., He, X., Wang, T., 2015. A consistent aerosol optical depth (AOD) dataset over mainland China by integration of several AOD products. Atmos. Environ. 114, 48–56.
Xue, T., Liu, J., Zhang, Q., Geng, G., Zheng, Y., Tong, D., Liu, Z., Guan, D., Bo, Y., Zhu, T., He, K., Hao, J., 2019a. Rapid improvement of PM2.5 pollution and associated health benefits in China during 2013–2017. Sci. China Earth Sci. 62, 1847–1856.
Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019b. Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000-2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357.
Yang, J., Hu, M., 2018. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. Sci. Total Environ. 633, 677–683.
Yu, C., Di Girolamo, L., Chen, L., Zhang, X., Liu, Y., 2015. Statistical evaluation of the feasibility of satellite-retrieved cloud parameters as indicators of PM 2.5 levels. J. Expo. Sci. Environ. Epidemiol. 25, 457–466.
Yue, H., He, C., Huang, Q., Yin, D., Bryan, B.A., 2020. Stronger policy required to substantially reduce deaths from PM2. 5 pollution in China. Nat. Commun. 11.
Zhang, Z., Wu, W., Fan, M., Wei, J., Tan, Y., Wang, Q., 2019. Evaluation of MAIAC aerosol retrievals over China. Atmos. Environ. 202, 8–16.
Zhao, C., Liu, Z., Wang, Q., Ban, J., Chen, N.X., Li, T., 2019. High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing-Tianjin-Hebei region. Atmos. Environ. 203, 70–78.
Zhao, Y., Huang, B., Liu, D., He, Q., 2021. A sparse representation-based fusion model for improving daily MODIS C6. 1 aerosol products on a 3 km grid. Int. J. Remote Sens. 42, 1077–1095.
Zheng, Q., Weng, Q., Wang, K., 2021. Characterizing urban land changes of 30 global megacities using nighttime light time series stacks. ISPRS J. Photogramm. Remote Sens. 173, 10–23.