https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


# LGHAP: a Long-term Gap-free High-resolution Air Pollutants concentration dataset derived via tensor flow based multimodal data fusion

Kaixu Bai[1,2*], Ke Li[1], Mingliang Ma[3], Kaitao Li[4], Zhengqiang Li[4], Jianping Guo[5*], Ni-Bin Chang[6], Zhuo Tan[1], Di Han[1]

[1]Key Laboratory of Geographic Information Science (Ministry of Education), School of Geographic Sciences, East China Normal University, Shanghai 200241, China

[2]Institute of Eco-Chongming, 20 Cuiniao Rd., Chongming, Shanghai 202162, China

[3]School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China

[4]State Environmental Protection Key Laboratory of Satellite Remote Sensing, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China

[5]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

[6]Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL, USA

[*]**Correspondence to:** Kaixu Bai (kxbai@geo.ecnu.edu.cn) and Jianping Guo (jpguocams@gmail.com)

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Earth System Science Data
Open Access Discussions

**Abstract**

Developing a big data analytics framework for generating a Long-term Gap-free High-resolution Air Pollutants concentration dataset (abbreviated as LGHAP) is of great significance for environmental management and earth system science analysis. By synergistically integrating multimodal aerosol data acquired from diverse sources via a tensor flow based data fusion method, a gap-free aerosol optical depth (AOD) dataset with daily 1-km resolution covering the period of 2000–2020 in China was generated. Specifically, data gaps in daily AOD imageries from MODIS aboard Terra were reconstructed based on a set of AOD data tensors acquired from satellites, numerical analysis, and *in situ* air quality data via integrative efforts of spatial pattern recognition for high dimensional gridded image analysis and knowledge transfer in statistical data mining. To our knowledge, this is the first long-term gap-free high resolution AOD dataset in China, from which spatially contiguous $PM_{2.5}$ and $PM_{10}$ concentrations were estimated using an ensemble learning approach. Ground validation results indicate that the LGHAP AOD data are in a good agreement with *in situ* AOD observations from AERONET, with R of 0.91 and RMSE equaling to 0.21. Meanwhile, $PM_{2.5}$ and $PM_{10}$ estimations also agreed well with ground measurements, with R of 0.95 and 0.94 and RMSE of 12.03 and 19.56 μg $m^{-3}$, respectively. Overall, the LGHAP provides a suite of long-term gap free gridded maps with high-resolution to better examine aerosol changes in China over the past two decades, from which three distinct variation periods of haze pollution were revealed in China. Additionally, the proportion of population exposed to unhealthy $PM_{2.5}$ was increased from 50.60% in 2000 to 63.81% in 2014 across China, which was then drastically reduced to 34.03% in 2020. Overall, the generated LGHAP aerosol dataset has a great potential to trigger multidisciplinary applications in earth observations, climate change, public health, ecosystem assessment, and environmental management. The daily resolution AOD, $PM_{2.5}$, and $PM_{10}$ datasets can be publicly accessed at https://doi.org/10.5281/zenodo.5652257 (Bai et al., 2021a), https://doi.org/10.5281/zenodo.5652265 (Bai et al., 2021b), and https://doi.org/10.5281/zenodo.5652263 (Bai et al., 2021c), respectively. Meanwhile, monthly and annual mean datasets can be found at https://doi.org/10.5281/zenodo.5655797 (Bai et al., 2021d) and https://doi.org/10.5281/zenodo.5655807 (Bai et al., 2021e), respectively. Python, Matlab, R, and IDL codes were also provided to help users read and visualize these data.

**Keywords:** Aerosol optical depth; Particulate matter; Gap filling; Big data analytics; Multimodal data fusion

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Open Access Earth System Science Data Discussions

## 1 Introduction

Atmospheric aerosols not only impact regional climate by changing the Earth radiation budget but significantly influence air quality at the ground level (Fuzzi et al., 2015; Gao et al., 2018; Shen et al., 2020; Sun et al., 2015). Monitoring aerosol loading in the atmosphere is thus of great significance for climate change attribution and haze pollution assessment. Aerosol optical depth (AOD), a measure of aerosols distributed within a column of air from the Earth's surface to the top of the atmosphere, has been monitored for decades to quantify aerosol loading in the atmosphere. Compared with sparsely distributed ground aerosol monitoring stations (e.g., AERONET), satellite instruments can provide better AOD observations because of vast spatial coverage and high sampling frequency. An overview of sensors, algorithms, and AOD datasets that are widely used can be found in the literature such as Sogacheva et al. (2020) and Wei et al. (2020).

Due to negative impacts of bright surface (e.g., snow cover) and clouds, as well as algorithmic restrictions, satellite AOD retrievals often suffer from extensive data gaps, significantly reducing the downstream application potential such as mapping particulate matter (PM) concentrations at the ground surface (e.g., Bai et al., 2019; Wei et al., 2021a). Also, data gaps in AOD imageries may result in large uncertainty when assessing aerosol impacts on weather and climate (Guo et al., 2017; Li et al., 2019; Zhao et al., 2020). Over the years, many gap filling methods have been developed (e.g., Bai et al., 2016, 2020b; Chang et al., 2015). Nonetheless, filling data gaps in satellite-based AOD products is still a challenge due to extraordinary nonrandom missing values and high aerosol dynamics in space and time.

Wei et al. (2020a) provided a short review of methods that have been frequently applied to deal with data gaps in AOD products. In general, merging AOD data acquired from diverse instruments and/or platforms is the most popular approach to improve AOD spatial coverage (Sogacheva et al., 2020). Statistical methods such as linear regression (Bai et al., 2019a; Wang et al., 2019; Zhang et al., 2017), inversed variance weighting (Chen et al., 2018; Ma et al., 2016; Sogacheva et al., 2020), and maximum likelihood estimate (Xu et al., 2015), are often applied to account for systematic bias among different datasets. Data fusion methods such as Bayesian maximum entropy could be applied to blend AOD products with different resolutions (Tang et al., 2016; Wei et al., 2021b). Another way is to reconstruct missing AOD values using either neighboring observations in space and time or external

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




data sources such as AOD simulations from numerical models (Li et al., 2020; Xiao et al., 2017a), even simply meteorological factors (Bi et al., 2018).

Although there exist many versatile gap filling methods, spatially gap free AOD datasets are always rare, particularly satellite-based high-resolution AOD datasets, resulting in significant limit in downstream applications such as $PM_x$ concentration mapping. In spite of versatile $PM_{2.5}$ concentration prediction models (e.g., Di et al., 2019; Fang et al., 2016; He et al., 2020; Hu et al., 2014; Li et al., 2018b, 2016; Lin et al., 2016; Liu et al., 2009; Ma et al., 2014; Wang et al., 2021a), to date, there are few publicly accessible PM concentration datasets that can be used to examine haze pollution variations regionally and globally. Aa typical datasets, the one generated by the Dalhousie University (van Donkelaar et al., 2010, 2016), CHAP (Wei et al., 2019a), and TAP (Geng et al., 2021a) demonstrated the global effort to elevate earth system science research. However, these datasets more or less suffer from drawbacks in spatial and/or temporal resolution, spatial coverage, and data accuracy. To meet the contemporary needs, Zhang et al. (2021) provided a more comprehensive review of the widely used PM concentration mapping approaches.

With a thorough review for $PM_{2.5}$ concentration mapping techniques, an optimal full-coverage $PM_{2.5}$ concentration mapping scheme was proposed, in which diverse aerosol datasets were fused toward a full-coverage AOD map based on a multi-modal approach (Bai et al., 2021). In parallel with these efforts, some attempted to improve AOD data coverage over space with high accuracy by merging AODs observed at adjacent times directly (Li et al., 2021). Given such prior knowledge, the current study developed a big data analytics framework for generating a Long-term Gap-free High-resolution Air Pollutants concentration dataset (abbreviated as LGHAP hereafter) providing AOD, $PM_{2.5}$ and $PM_{10}$ concentration with a daily 1-km resolution in China from 2000 to 2020. Multimodal aerosol data acquired from diverse sources including satellites, ground stations and numerical models were synergistically integrated via the higher order singular value decomposition (HOSVD) to form a tensor flow based data fusion method in the current study. Full coverage $PM_{2.5}$ and $PM_{10}$ concentration data were then estimated on the basis of the gap-filled AOD dataset. This 21-year-long gap-free high resolution (daily/1km) aerosol dataset was then compared against ground-based AOD and PM observations to evaluate the data accuracy of each product, particularly the performance in spatial pattern recognition and temporal trend assessment. These advances led to explore the long-term

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


variability and population exposure to haze pollution in China over the past two decades by taking advantage of the LGHAP dataset.

## 2 Data sources

Table 1 summarizes the multisource datasets used in this study to help generate the LGHAP product. As shown, six satellite-based AOD products, five numerical simulations of AOD and aerosol components, eleven meteorological factors, six ground-based AOD and air quality datasets, as well as five land cover, topographic and socioeconomic parameters, were employed. Descriptions of these datasets are given in the following subsections.

**Table 1.** Summary of the data sources used in this study to generate gap free high resolution AOD and PMx concentration datasets.

| Category | Source product | Time range | Temporal resolution | Spatial resolution |
|---|---|---|---|---|
| AOD | Terra/MODIS | 2000–2020 | daily | 1 km |
| | Aqua/MODIS | 2002–2020 | daily | 1 km |
| | Terra/MISR | 2000–2020 | daily | 4.4 km |
| | Suomi-NPP/VIIRS | 2012–2020 | daily | 5 km |
| | Envisat/AATSR | 2000–2012 | daily | 10 km |
| | PARASOL/POLDER | 2005–2013 | daily | 10 km |
| | MERRA-2 | 2000–2020 | hourly | 0.5°×0.625° |
| | AERONET | 2000–2020 | hourly | point |
| Meteorology | Air temperature | 2000–2020 | hourly | 0.25° |
| | U/V component of wind | | hourly | 0.25° |
| | Relative humidity | | hourly | 0.25° |
| | Surface pressure | | hourly | 0.25° |
| | Boundary layer height | | hourly | 0.25° |
| | Total column water vapor | | hourly | 0.25° |
| | Surface solar radiation downwards | | hourly | 0.25° |
| | Instantaneous moisture flux | | hourly | 0.25° |
| | Visibility | 2000–2013 | 3-hour | point |
| Air quality | $PM_{2.5}$, $PM_{10}$, $SO_2$, $NO_2$ | 2014–2020 | hourly | point |
| Population | WorldPop | 2000–2020 | annual | 1 km |
| Elevation | DEM | 2000 | / | 30 m |
| Land Cover | CLCD | 2000–2019 | annual | 30 m |
| | GLOBELAND | 2020 | annual | 30 m |
| NDVI | Terra/MODIS | 2000–2020 | monthly | 1 km |
| Aerosol component | MERRA-2 | 2000–2020 | hourly | 0.5°×0.625° |

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



## 2.1 Gridded aerosol products

In many previous studies, coarse AOD and/or aerosol components simulations acquired from numerical models were oftentimes used as the primary data source to help derive full-coverage AOD and/or $PM_{2.5}$ concentration maps (e.g., Park et al., 2020; Wang et al., 2021b). However, due to the lack of high accuracy near real-time emission inventory, simulated AOD and/or aerosol components are often prone to large uncertainty, which could be inevitably introduced to the final $PM_{2.5}$ estimations if no observational data are applied for bias correction. In such a research context, here we used six satellite-based AOD products with a relatively long-term coverage to help better reconstruct historical AOD variations over space and time.

The latest AOD product derived from the MODerate-resolution Imaging Spectroradiometer (MODIS) onboard Terra using the multiangle implementation of atmospheric correction (MAIAC) algorithm (Lyapustin et al., 2011, 2018), was hereby used as the baseline dataset for the generation of full-coverage AOD maps. This AOD product has not only a finer spatial resolution but a comparable and even better accuracy, when comparing with those derived from the Dark Target and Deep Blue algorithms (Goldberg et al., 2019; Lyapustin et al., 2018; Xiao et al., 2017b). In addition, AOD products derived from MODIS onboard Aqua, the Multi-angle Imaging SpectroRadiometer (MISR) onboard Terra, Visible Infrared Imaging Radiometer Suite (VIIRS) onboard Suomi-NPP, Advanced Along-Track Scanning Radiometer (AATSR) onboard Envisat and POLarization and Directionality of the Earth's Reflectances (POLDER) onboard PARASOL, were also utilized. The ultimate goal was to reduce bias in full-coverage AOD imagery by better spatial coverage of observational AOD as much as possible. Accuracies of these AOD products have been extensively validated, e.g., de Leeuw et al. (2018), Xiao et al. (2016), Wei et al. (2019b), Che et al. (2019), to name a few, in the current study. A brief description of these satellite-based AOD products can be found in Text S1 in the supplementary information.

In addition to satellite-based AOD products, numerically simulated aerosol diagnostics from MERRA-2, including AOD and aerosol components such as black carbon, organic carbon, dust and sulfate, were also applied to help reconstruct missing AOD information and to predict $PM_{2.5}$ and $PM_{10}$ concentrations at the ground level. The aerosol components were used here as a proxy of emission inventory when predicting PM concentrations. Big data analytics procedures applied to these datasets will be described in section 3.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




## 2.2 *In situ* AOD and air quality measurements

AOD observations from Aerosol Robotic Network (AERONET) were used to evaluate the prediction accuracy of the generated full-coverage (gap free) AOD product, as well as the learning target to infer AOD from air pollutants concentration and atmospheric visibility. Considering few valid data were provided in the Level 2.0 dataset, here we used the Level 1.5 AOD data to guarantee adequate *in situ* AOD data coverage in space and time. To validate the gridded AOD products in this study, each *in situ* AOD observation was registered with the gridded mean AOD over a 50×50 km window.

Near-surface air pollutants concentrations including $PM_{2.5}$, $PM_{10}$, $NO_2$, and $SO_2$ that were sampled at state-controlled monitoring sites were also applied, not only to help establish machine-learned regression models for PM prediction ($PM_{2.5}$ and $PM_{10}$), but to infer AOD over air quality monitoring sites given their dense distribution across China. The gauged air pollutants concentration data have been released online on an hourly basis by the China National Environment Monitoring Center since the late 2013. For quality control, outliers were first detected and removed from each pollutant dataset by following the criteria used in our previous study (Bai et al., 2020a). The missing values were then reconstructed using the diurnal cycle constrained empirical orthogonal function (DCCEOF) method proposed in Bai et al. (2020b).

The 3-hour resolution atmospheric visibility data acquired from 4,052 weather stations were employed to help generate gap free AOD maps before 2014, at which *in situ* air quality measurements were not available. Previous studies have attempted to predict $PM_{2.5}$ concentration from atmospheric visibility data with good accuracies (Liu et al., 2017), indicative of a great potential for estimating AOD. Specifically, visibility data were used as an important predictor for site-specific AOD prediction, and the resulting AOD predictions were then used as a critical prior information for reconstructing AOD distributions over space, especially over those regions without satellite AOD observations. Since automatic visibility sensors have been widely used at many sites since 2014, those data were excluded to guarantee the data consistency (Li et al., 2018a). For quality control, the consistency of visibility data was examined using an outlier detection method, i.e., the annual mean should not exceed 3 times the standard deviation of data over a 5-year time window (Zhang et al., 2020). Those with apparent jumps and drifts in visibility time series were excluded. Meanwhile, visibility data on rainstorm and foggy days were eliminated as well.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


### 2.3 Auxiliary data

As shown in Table 1, eleven meteorological factors, including air temperature at the near surface, wind speed and direction, relative humidity, surface pressure, boundary layer height, total column water vapor, downwards solar radiation, and instantaneous moisture flux, were used to help infer $PM_{2.5}$ and $PM_{10}$ from AOD, as well as to downscale AOD from MERRA-2. These data were acquired from the fifth generation ECMWF atmospheric reanalysis (ERA-5), and the first three factors were extracted at the levels of not only ground surface but 850 hpa and 500 hpa so as to indicate the vertical structure of the atmosphere. Additionally, population data from WorldPop, land cover from CLCD during 2000 to 2019 (Yang and Huang, 2021) and GLOBELAND 30 in 2020 (Chen et al., 2014), elevation data from the Global Digital Elevation Model (GDEM) version 2, as well as monthly composited 1-km normalized difference vegetation index (NDVI) from MODIS, were employed to indicate the socioeconomic and ecological contributions to haze pollutions. Properties of these datasets can be found in Table 1, and datasets with a finer resolution were upscaled to 0.01° via a cubic interpolation method.

## 3 Methodology

To advance environment management and earth system science analysis, the current study developed a big data analytics framework for generating long-term gap free aerosol spatiotemporal datasets and demonstrated its applications in China. Such big data analytics was constructed via a seamless integration of the tensor flow based multimodal data fusion with ensemble learning based PM concentration estimation. When generating this dataset, the proposed method transformed a set of data tensors of AOD and other related datasets such as air pollutants concentration and atmospheric visibility that were acquired from diversified sensors or platforms via integrative efforts of spatial pattern recognition for high dimensional gridded data analysis toward data fusion and multiresolution image analysis, as well as knowledge transfer in statistical data mining. The proposed method consists of three major procedures in general, including multisensory data homogenization, tensor flow based AOD reconstruction, and ensemble learning for PM concentration estimation. The analytical framework of the big data analytics is depicted in Figure 1 and described in details in the following subsections.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



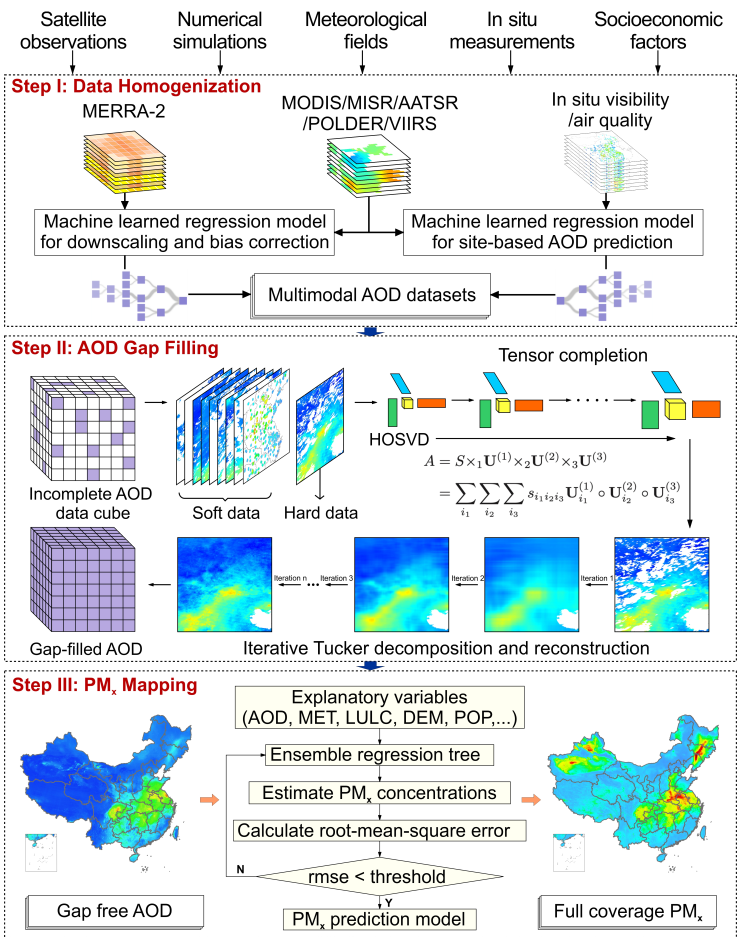




**Figure 1.** Flowchart of the proposed big data analytics framework for generating a long-term gap-free high-resolution air pollutants concentration dataset (LGHAP), taking AOD and PM concentration in China as illustration.

### 3.1 Multisensory data homogenization

Since a set of aerosol products with different types, resolution, and accuracies were applied to support the generation of gap-free AOD imageries, harmonizing cross-mission biases and scale differences between these diversified datasets is thus of critical importance to facilitate multisensory data integration. In this study, machine-learned regression models were established to harmonize these heterogeneous aerosol datasets. A baseline dataset was first selected to be used as the learning target while other datasets were calibrated to the level of baseline dataset to make them comparable. Given

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



finer resolution and higher proportion of data coverage in space and time, the MAIAC AOD product from Terra ($AOD_{Terra}$) was selected as the baseline dataset. Consequently, six machine-learned regression models were established between $AOD_{Terra}$ and each gridded AOD product (i.e., five satellite-based AOD products plus MERRA-2 AOD simulations) using the random forest method. Meteorological factors, land cover, topographic and socioeconomic variables were used as covariates to help downscale these multimodal AOD products to have a resolution same as $AOD_{Terra}$ while accounting for cross-mission biases arising from temporal and algorithmic differences.

Considering data gaps are extensive in satellite-based AOD products, especially over regions with thick cloud cover, providing prior AOD information over such region is thus of great value in support of the reconstruction of missing AOD values. As indicated in our recent studies, AOD can be accurately predicted from ground measured air pollutants concentration, showing an accuracy even over some satellite AOD retrievals (Li et al., 2021; Bai et al., 2021). To support AOD reconstruction over regions without satellite AOD observations, we attempted to infer AOD over air quality monitoring sites from air pollutants concentration measurements via an ensemble learning approach. Similarly, machine-learned regression models were established using random forest by taking $AOD_{Terra}$ as the learning target while ground measured air pollutants concentration, meteorological factors, land cover, and terrain information, were used as predictors.

The transformation of ground measured air pollutants concentration data to AOD empowers us to provide external observational AOD to supplement satellite observations, especially over regions suffering from significant data gaps. Since air pollutants concentration data were not available before 2013, atmospheric visibility data sampled at dense weather stations were hereby used as an alternative for AOD prediction, by applying a similar prediction model as used above for air pollutants concentration. Figure S1 show the ground-based validation results of AOD inferred from atmospheric visibility and air pollutants concentration, indicative of a generally good accuracy of these inferred AOD values. All efforts led to aggregate a set of multimodal aerosol data with different properties for multisensory data fusion toward generating full-coverage (gap free) AOD mapping as the next step.

## 3.2 Tensor flow based AOD reconstruction

The core of generating full coverage AOD imageries is to fill in data gaps in $AOD_{Terra}$. Previous studies have demonstrated that merging satellite AOD retrievals derived at adjacent time steps can

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



help improve the observational AOD coverage at each single snapshot, while the involvement of numerical AOD simulations can help fill in AOD data gaps (Li et al., 2021; Bai et al., 2021). In this study, a tensor completion method was particularly designed and applied to fulfil the gap filling in $AOD_{Terra}$. Specifically, the incomplete $AOD_{Terra}$ imageries were deemed as the hard data (true AOD state) while other AODs datasets (e.g., the downscaled AOD datasets and site-specific AOD predictions inferred from air pollutants concentration and atmospheric visibility) were used as the soft data to help reconstruct AOD distribution in $AOD_{Terra}$ via tensor flow based pattern recognition. Detailed procedures for gap filling are outlined as follows.

### 3.2.1 Initial AOD tensor construction

Due to extensive data gaps in satellite-based AOD retrievals, it is insufficient to reconstruct all missing AOD information in $AOD_{Terra}$ for a given date by simply using the harmonized satellite-based AOD data synchronously. To fulfill AOD gap filling, the newly developed tensor completion method was thus applied to synergistically integrate AOD acquired from diversified sources. Consequently, creating the data tensor of AOD is of critical importance. In this study, the data tensor of AOD was constructed by incorporating not only observational AOD from both satellites and those inferred from *in situ* air quality indicators on the same date, but also historical AOD observations from MODIS instruments and part of data from the downscaled MERRA-2 AOD (denoted as $AOD_{M2}$ hereafter). The latter two were applied to provide knowledge of AOD distributions over space to guide the reconstruction of missing values in $AOD_{Terra}$.

For the screening of historical observations resembling $AOD_{Terra}$ distribution on the given date, $AOD_{M2}$ was used in concert with $AOD_{Terra}$ and site-based AOD estimations to select similar imageries. Specifically, site-specific AOD estimations and 5% randomly selected $AOD_{M2}$ data were merged with valid $AOD_{Terra}$ to form a new image on each date, which was then used to find similar historical $AOD_{Terra}$ maps. To avoid the inclusion of imageries with distinct variation patterns, only those closely resembling AOD distribution in the composite image on the given date were selected in terms of their correlations and biases subject to a threshold of R>0.7 and RMSE<0.2. Once sufficient historical imageries were obtained, the data tensor of AOD was constructed by compiling the observed AOD imageries on the given date with historical imageries to a three-dimension data array $\mathbf{A} \in \mathbf{R}^{N_1 \times N_2 \times N_3}$ (composed of $N_3$ images with a size of $N_1 \times N_2$). Data values of site-specific AOD estimations and 1%

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



randomly selected $AOD_{M2}$ data were directly placed on grids where $AOD_{Terra}$ values missed on each specific date. This greatly facilitated the reconstruction of missing AOD information over regions with tremendous data gaps in satellite observed AOD imageries given the presence of prior knowledge. More importantly, it significantly reduced the time required for convergence during the gap filling process.

### 3.2.2 Gap filling via tensor completion

Previous studies have well demonstrated the good performance of matrix decomposition methods such as empirical orthogonal function and singular value decomposition (SVD) for missing value imputation (Bai et al., 2020b; Beckers and Rixen, 2003; Folch-Fortuny et al., 2015). However, these methods can only work on two-dimension matrix mathematically, i.e., the matrix domain. To integrate spatial features of AOD revealed by datasets to generate a smooth AOD distribution with complete coverage, in this study, the HOSVD, a specific orthogonal Tucker decomposition, was applied. More detailed descriptions to HOSVD can be found in the literature such as Sun et al. (2021), Tucker (1966), Kolda and Bader (2009), Sidiropoulos et al. (2017), and Chen et al. (2014).

In Table 2, we provided a stepwise description of the algorithm used to fill data gaps in $AOD_{Terra}$ by integrating AOD features recognized in different imageries as the data tensor of AOD via HOSVD. To initiate the tensor decomposition, grids with missing values in the AOD tensor were first filled with the spatial average of valid AOD data in each individual image. Then, the AOD tensor was decomposed along each of three dimensions, while the dominant features in each dimension determined by the corresponding rank values were applied to reconstruct the data tensor. By gradually increasing the rank values and iteratively updating the initial filled values, the tensor can be reconstructed to better delineate AOD distribution over space after several iterations.

To confirm the convergence, a small portion of observational AOD values were randomly held out in advance, and the reconstructed values over these grids in each iteration were compared with these hold-out data till the difference between them lower than 0.01 (a threshold to determine convergence, a.k.a, $\varepsilon_1$ in Table 2). Meanwhile, to make the computational burden manageable, the study region (China in this study) was divided into 40 subregions (refer to Figure S2 for the spatial distribution of these subregions), and the tensor completion was then performed over each individual region. Finally, the reconstructed imageries were mosaiced to attain a national gap-free AOD map on each specific

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




date. During this step, a weighted average method was applied to solve the boundary effect when mosaicking two adjacent maps, i.e., averaging the data value on each overlapped grid at the boundary (50 km on the edge of subregion) as weighted by the distance to the edge. In the end, the mosaic $AOD_{Terra}$ image was retained as the final gap-free AOD product.

**Table 2.** The proposed tensor completion algorithm for AOD distribution reconstruction in $AOD_{Terra}$.

**Input:** tensor $\mathbf{A} \in \mathbf{R}^{N_1 \times N_2 \times N_3}$ with $\boldsymbol{\Omega} = \{(\mathrm{i,j,k}): A_{ijk} \text{ is observed}\}$, threshold $\mathrm{T}_1$, $\mathrm{T}_2$
**Output:** reconstructed entries $\mathbf{A}' = \mathbf{A}^*(:,:,k^t) \in \mathbf{R}^{N_1 \times N_2}$

Initialize $\mathrm{A}^*_{ijk} = \begin{cases} A_{ijk} & (i,j,k) \in \boldsymbol{\Omega} \\ \sum_i \sum_j A_{ijk} & (i,j,k) \notin \boldsymbol{\Omega} \end{cases}$
**for** $\mathrm{n}_3 = N_3$ to 1 **do**
  $\mathrm{n}_1 = n_2 = 0$
  **while** $\varepsilon_1 > T_1$ **do**
    $n_1 = n_1 + 1, n_2 = n_2 + 1$
    Tucker Decomposition of $\mathbf{A}^*$ with rank $= \{\mathrm{n}_1, \mathrm{n}_2, \mathrm{n}_3\}$:
    $\mathbf{A}^* = S \times_1 \boldsymbol{U}^{(n_1)} \times_2 \boldsymbol{U}^{(n_2)} \times_3 \boldsymbol{U}^{(n_3)}$
    $\varepsilon_1 = \arg\min_{\boldsymbol{\Omega}} \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{A}^*\|^2$
    $\boldsymbol{A}^*_{\boldsymbol{\Omega}} = \boldsymbol{A}_{\boldsymbol{\Omega}}$
  **end while**
  **if** $\arg\min_{\boldsymbol{\Omega}} \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{A}^*\|^2 < T_2$ **then**
    break;
  **end if**
**end for**

## 3.3 PM concentration estimation

In this study, the random forest method was applied to establish regression models for $PM_{2.5}$ and $PM_{10}$ concentration mapping. Ground measured $PM_{2.5}$ (or $PM_{10}$) concentration data were used as the learning target while AOD, aerosol components ($AER_{comp}$), meteorological factors (MET), digital elevation model (DEM), NDVI, land cover information (LC), and population were used as regressors. The prediction model can be generally formulated as:

$$\mathrm{PM_x} = \mathrm{f(AOD, AER_{comp}, MET, DEM, NDVI, POP, LC, month)} \quad (1)$$

where month is a categorical variable that was used to account for monthly varying relationships between AOD and PM. For cross validation, $PM_{2.5}$ and $PM_{10}$ data from 10% of monitoring sites were randomly held out to validate the predictive performance of each regression model. 500 regression trees were used in each RF model, and each tree was grown on a bootstrap sample. The learning data set was randomly divided into two parts during the training process, with 80% used as the training set while the rest 20% for testing. In order to guarantee a larger value of $PM_{10}$ than $PM_{2.5}$, $PM_{2.5}$ estimations from Eq. (1) were used as one predictor in addition to factors used to predict $PM_{2.5}$ when estimating $PM_{10}$ concentration. Such a model can also significantly improve the prediction accuracy of $PM_{10}$ given the prior $PM_{2.5}$ information.

### 3.4 Point-surface data fusion

Ground measured $PM_{2.5}$ and $PM_{10}$ concentration data were further fused with gridded PM estimations to enhance the data accuracy of PM data after 2014. Here, the well-known optimal interpolation (OI) method was applied to perform point-surface fusions between two different types datasets. Please refer to Bai et al. (2021) and Li et al. (2021) for a more detailed description of the OI method used to fuse PM concentration data. In this study, a modified scheme was developed to select neighboring observations. To avoid an isotropic interpolation effect, here we only used 30 ground observations with land cover, terrain and atmospheric conditions similar to those at the analyzed grid cell to estimate the innovation that should be assigned to the background value at the given grid. In other words, a similarity measure was first estimated between the analyzed grid cell and neighboring sites in terms of land cover, DEM, and atmospheric conditions. The 30 observations with similar background fields were then used in the OI procedure to correct possible bias in gridded PM estimations. Such a treatment can help exclude those observations with different ambient background, e.g., one site not far from the given grid but separated by a high mountain, thereby avoiding the possible propagation of antiphase corrections to data over adjacent grids.

## 4 Results and discussion

### 4.1 Data accuracy of gap-free AOD in LGHAP

Table 3 summarizes the data accuracy of gap-free AOD dataset generated in this study. For comparison, the data accuracy of each original AOD dataset was also assessed. Since *in situ* AOD measurements were not used as data input when reconstructing missing AOD information, thereby the gap-free AOD can be directly compared with *in situ* AOD measurements. As indicated, all these AOD datasets are in a good agreement with *in situ* AOD measurements. Generally, AODs from MODIS onboard Terra and Aqua have a similar data accuracy, which is also among the highest when

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


comparing with other datasets (R=0.95 and RMSE=0.14). AODs from AATSR show a comparable accuracy with that of MODIS, but with a relatively low correlation with ground AOD measurements. AODs from MISR, POLDER and VIIRS exhibit a similar bias level, with R varying from 0.80 to 0.92 and RMSE ranging from 0.22 to 0.29. In contrast, $AOD_{M2}$ data have the poorest accuracy among these eight gridded AOD datasets (R=0.77 and RMSE=0.36), even though AOD data from AERONET and satellite observations like MODIS had been already assimilated. This indicates the presence of large biases in $AOD_{M2}$ and thus these $AOD_{M2}$ data cannot be used solely to delineate AOD distributions over space.

**Table 3.** Data accuracy of original and gap-free AOD datasets used and/or generated in this study. The expected error (EE) was defined as $\pm 0.05 + 0.15 \times AOD_{site}$.

| Dataset | N | R | RMSE | MAE | Below EE (%) | Within EE (%) | Above EE (%) |
|---|---|---|---|---|---|---|---|
| Terra/MODIS | 6731 | 0.95 | 0.13 | 0.07 | 8.94 | 78.73 | 12.33 |
| Aqua/MODIS | 6079 | 0.95 | 0.14 | 0.08 | 8.24 | 79.45 | 12.30 |
| Terra/MISR | 638 | 0.90 | 0.29 | 0.13 | 21.63 | 73.51 | 4.86 |
| NPP/VIIRS | 3839 | 0.80 | 0.22 | 0.16 | 7.03 | 44.93 | 48.03 |
| Envisat/AATSR | 434 | 0.92 | 0.11 | 0.07 | 17.74 | 73.96 | 8.29 |
| PARASOL/POLDER | 1733 | 0.92 | 0.24 | 0.17 | 5.14 | 40.22 | 54.65 |
| MERRA-2 | 22067 | 0.77 | 0.36 | 0.20 | 32.97 | 51.76 | 15.27 |
| LGHAP | 24861 | 0.91 | 0.21 | 0.13 | 12.27 | 59.00 | 28.73 |

Compared to the first seven gridded AOD datasets, the LGHAP AOD dataset has an accuracy slightly worse than the original MODIS AOD product but comparable to AODs from MISR, POLDER and MERRA-2, with R of 0.91 and RMSE equaling to 0.21 compared to ground-based AOD observations. Compared with AODs from MODIS, the gap-filled AOD appeared to overestimate ground-based AOD observations, and this could be caused by the involvement of AODs from VIIRS and POLDER as these two products of ground AOD observations significantly overestimated, which can be indicated by the proportion of data pairs above the expected error (EE). On the other hand, such significant underestimations in $AOD_{M2}$ were not introduced to the LGHAP AOD as the former had a below EE ratio of 32.97% which was only12.27% in the latter. These results indicate that the gap-free LGHAP AOD data are more likely to resemble AOD distributions revealed by satellite observations

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




rather than $AOD_{M2}$, justifying the advantages of involving multiple satellite AOD observations to help reconstruct missing AOD values. Figure 2 further compares the data accuracy of $AOD_{Terra}$ and the reconstructed data over different regions of China. It is indicative that the purely reconstructed data have an accuracy (R=0.88 and RMSE=0.26) lower than the original $AOD_{Terra}$ (R=0.95 and RMSE=0.13) across China, especially in South China where the reconstructed data were significantly underestimated the ground-based AOD observations. Possible reasons for the relatively poor accuracy of AOD reconstructions in this region could be attributed to extensive data gaps over there due to frequent clouds (refer to Figure S3 for the distribution of mean data integrity of $AOD_{Terra}$ during 2000–2020), which significantly limit the learning capacity in space and temporal domain during the tensor completion process. In other words, limited observations in satellite imageries greatly reduced the learning performance from the sparse tensor. Even though, the purely reconstructed data exhibit a bias level comparable to AOD retrievals from several satellite instruments, e.g., MISR, VIIRS, and POLDER. This demonstrates the good performance of the proposed tensor completion method in reconstructing missing AOD information. By combining the reconstructed data with original AOD observations from Terra/MODIS, we obtained a 21-year-long spatially complete high-resolution (daily/1-km) AOD product with satisfying accuracy (R=0.91 and RMSE=0.21).

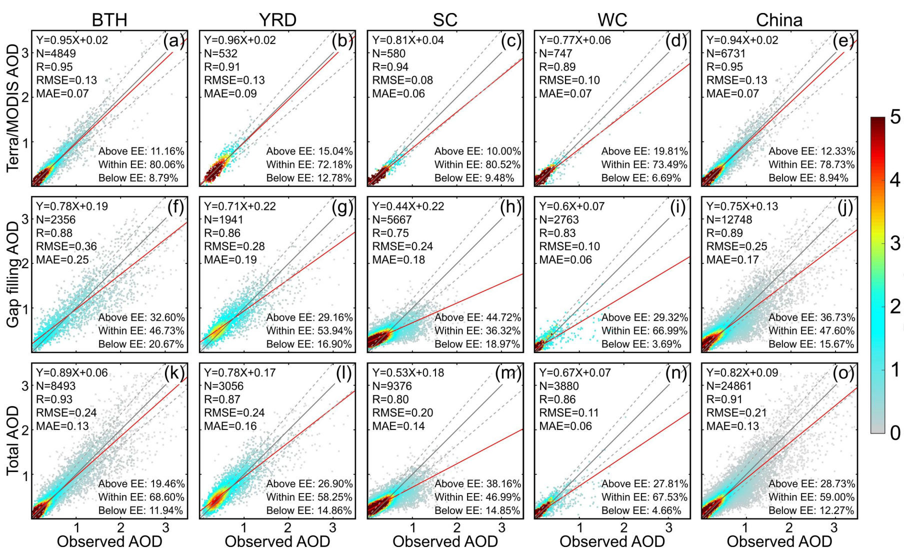



**Figure 2.** Scatter plots between ground observed and satellite-based AOD data in different regions of China. (a–e) original Terra/MODIS AOD, (f–j) reconstructed AOD only, and (k–o) both original and

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


reconstructed data combined. BTH, YRD, SC, and WC refers to regions of Beijing-Tianjin-Hebei, Yangtze River Delta, South China, and West China, respectively.

In Figure 3 we presented a comparison of AOD time series between the LGHAP dataset and ground observations at three AERONET sites under different air pollution levels. As shown, the AOD time series from LGHAP are temporally continuous whereas data gaps are common in AERONET observations. Generally, AODs from LGHAP are well reconstructed with respect to the temporal variations of aerosol loading at these three sites, with R ranging from 0.77 to 0.90 and RMSE varying between 0.11 and 0.21. For illustration, Figure 4 compares the spatial distribution of original and reconstructed AOD on four days with different $AOD_{Terra}$ coverage over space. As shown, the missing AOD values were well reconstructed after gap filling, resembling a smooth and reasonable AOD distribution over space, even over regions with very limited prior AOD observations from Terra/MODIS (e.g., Figure 4d). As indicated in Figures 4a and 4c, the high AOD loading was also properly reconstructed even though no prior information was provided by $AOD_{Terra}$. Since ground-based AOD observations were not used as a data input when generating the LGHAP AOD dataset, these independent validation results clearly demonstrated the high accuracy of the LGHAP AOD product as well as a good performance of the proposed full-coverage (gap free) AOD mapping.

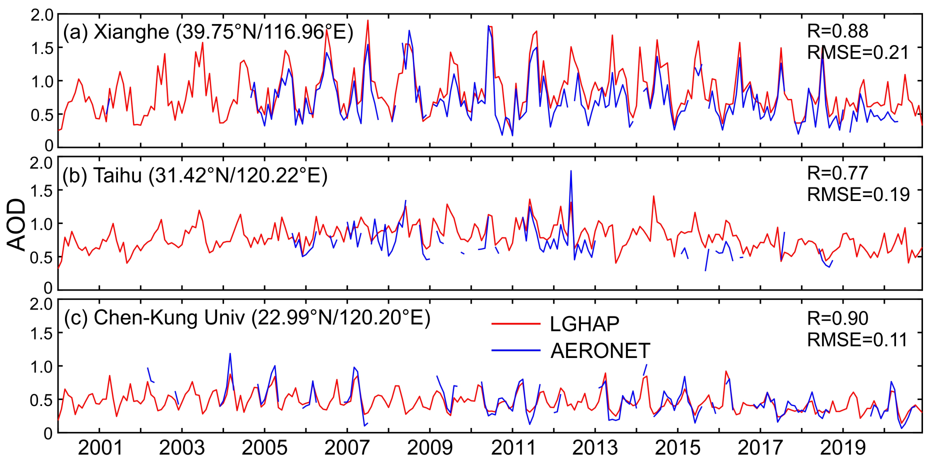



**Figure 3.** Comparison of monthly AOD time series from LGHAP and AERONET at three different stations in China. Latitude and longitude information of each site was given in brackets.

Since the final gap-free AOD product was generated mainly by integrating a set of data tensor of gridded AOD with *in situ* AOD estimations, the relative contribution of each product to the final

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


gap-free dataset is worth being investigated. In this study, a data coverage ratio weighted nonlinear correlation coefficient was proposed to examine the relative contribution of each gridded product to the LGHAP AOD dataset. The nonlinear correlation coefficient was used to assess the mutual information between two variables (Sun et al., 2021; Wang et al., 2005), while the data coverage ratio was multiplied to indicate the overall contribution of one product to the final fused dataset (refer to Text S2 for the definition of this indicator). As shown in Figure 5, the relative contribution of each gridded product varied with time and the input data sources. In the early two years (2000–2001), the AOD distribution in gap-free imageries was determined largely by $AOD_{Terra}$ (81%), whereas this ratio decreased to about 30% when many other products were involved, especially AOD from Aqua and PARASOL. With the advent of VIIRS and the loss of PARASOL after 2012, the relative contribution changed drastically as AOD from MODIS and VIIRS played the dominant roles in reconstructing AOD distribution. Note the relative contribution of $AOD_{M2}$ remained lower than 10%, indicative of the greater importance of satellite observations in generating the LGHAP AOD product.

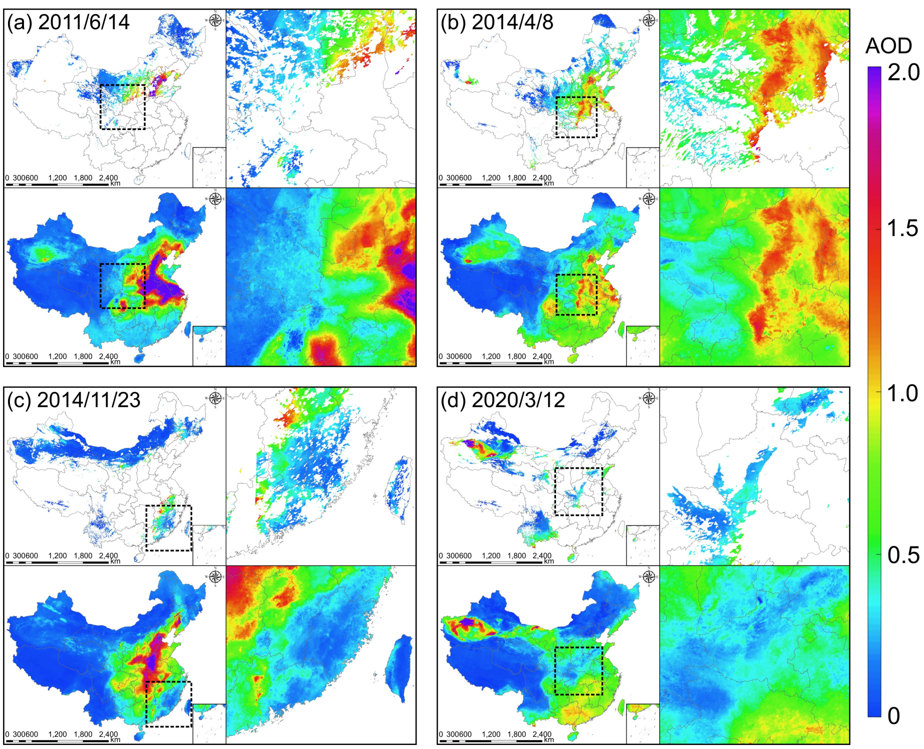



**Figure 4.** Spatial patterns of the reconstructed AOD under different baseline AOD coverage ratios. In each sub-diagram, the upper panel presents the original AOD distribution from Terra/MODIS while

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Open Access Earth System Science Data Discussions

the gap-filled imagery is shown below. The zoom-in views of the outlined regions are shown in the right part.

With respect to the temporally averaged contribution in each subregion, it shows that the relative contribution of each product also varied significantly across regions. Generally, AOD from MODIS aboard Terra and Aqua played the most important role (>60%) in generating the LGHAP AOD product, except over the southwest part of the country (Tibet plateau) where $AOD_{M2}$ contributed most. This is reasonable since data gaps are abnormally high in satellite observations over this region because of the vast and long-lasting snow cover (refer to Figure S3 for the data integrity distribution). Consequently, $AOD_{M2}$ would play an important role in reconstructing AOD distribution over such regions. Overall, the results shown here clearly highlight the success of big data analytics in generating the LGHAP AOD dataset via integrative efforts from diversified data sources.

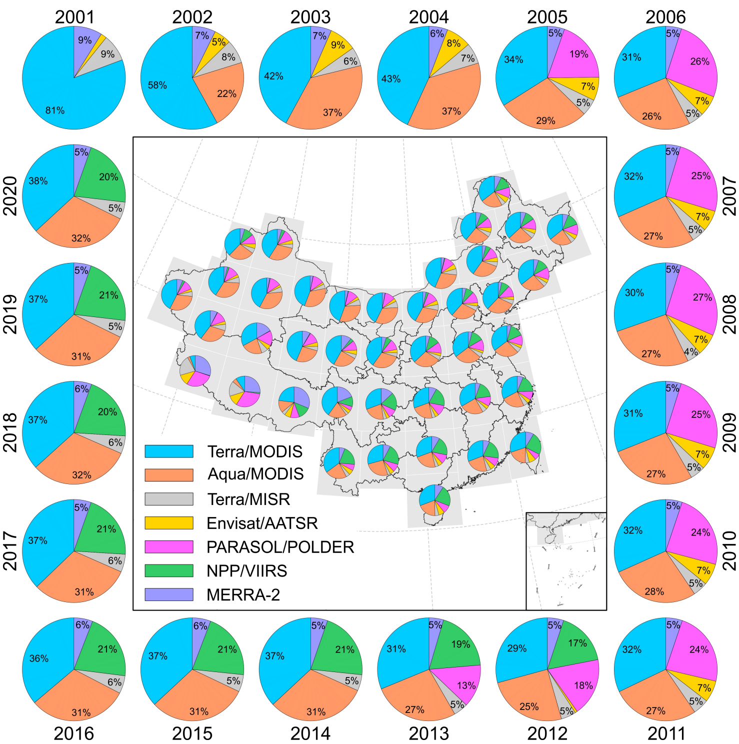



**Figure 5.** Spatiotemporal variations of the relative contribution of each gridded AOD product to the generation of LGHAP AOD dataset. The relative contribution was estimated as the data coverage ratio weighted nonlinear correlation coefficient (please refer to Text S2 in the supplementary information for the arithmetic theory to calculate this measure). The annual mean shown outside is the national

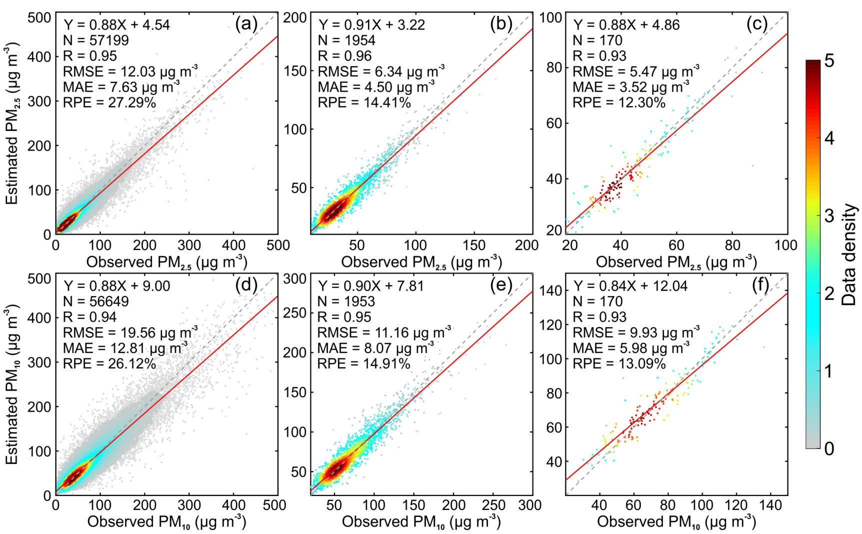
https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



averaged contribution in each individual year while the regional mean shown on the map was averaged over the past 21-year in each subregion.

### 4.2 Data accuracy of $PM_{2.5}$ and $PM_{10}$ estimations

By taking advantage of the gap-filled AOD, daily 1-km resolution $PM_{2.5}$ and $PM_{10}$ concentration data in China were estimated via an ensemble learning approach. Figure S4 shows the sample-based cross validation accuracy of two prediction models. It shows that the original daily $PM_{2.5}$ prediction model had a sample-based cross validation $R^2$ of 0.79 and RMSE of 20.04 μg $m^{-3}$. This accuracy is comparable to our previous study (Bai et al., 2019a), but slightly worse than those reported in some recent studies (Table 4). In contrast, $PM_{10}$ had a much higher prediction accuracy, with $R^2$ of 0.90 and RMSE of 21.06 μg $m^{-3}$ for the daily product. This good performance should be attributed to the involvement of $PM_{2.5}$ estimations as a predictor in the $PM_{10}$ prediction model. Figure 6 shows the site-specific (held-out in advance) validation accuracy of daily, monthly, and annual mean $PM_{2.5}$ and $PM_{10}$ concentration in LGHAP. As shown, the site-specific validation results indicated that the final full-coverage (gap free) daily $PM_{2.5}$ and $PM_{10}$ concentration data are in a good agreement with ground-based measurements, with R of 0.95 and RMSE of 12.03 μg $m^{-3}$ for $PM_{2.5}$ while R of 0.94 and RMSE of 19.56 μg $m^{-3}$ for $PM_{10}$. Overall, PM data in LGHAP are not only spatially complete with a finer resolution but have a comparable accuracy with previous studies.



https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




**Figure 6.** Scatter plots between observed and estimated $PM_{2.5}$ and $PM_{10}$ concentration. (a–c) respectively denotes daily, monthly, and annual mean $PM_{2.5}$ validation results, while (d–f) are for $PM_{10}$ concentration. The ground measurements were acquired from 30 independent air quality monitoring sites that were randomly held-out before the model training.

**Table 4.** Comparison of the data quality of $PM_{2.5}$ from LGHAP with other related studies.

| Source | Gap-free | Resolution | Time range | $R^2$ | RMSE (μg $m^{-3}$) |
|---|---|---|---|---|---|
| Wei et al. (2021) | No | 1 km | 2000~2018 | 0.86~0.90 | 10.09~18.39 |
| Geng et al. (2021) | Yes | 10 km | 2000~2021 | 0.80~0.88 | 13.90~22.10 |
| Xue et al. (2019) | Yes | 10 km | 2000~2016 | 0.61 | 27.80 |
| Chen et al. (2018) | No | 10 km | 2005~2016 | 0.83 | 28.10 |
| Lyu et al. (2019) | Yes | 12 km | 2014~2017 | 0.64 | 24.80 |
| Ma et al. (2016) | No | 10 km | 2004~2013 | 0.79 | 27.42 |
| Huang et al. (2021) | No | 1 km | 2013~2019 | 0.88 | 15.73 |
| Xiao et al. (2018) | Yes | 10 km | 2013~2017 | 0.79 | 21.00 |
| LGHAP $PM_{2.5}$ | Yes | 1 km | 2000~2020 | 0.90 | 12.03 |

Figure 7 presents a two-year-long comparison of $PM_{2.5}$ concentration time series from LGHAP and two other open access datasets with $PM_{2.5}$ measurements sampled at four United States Embassy in China. Since this ground-based dataset has been seldomly noticed and used, it can be applied as an independent dataset to fairly evaluate the accuracy of these three machine-learned $PM_{2.5}$ estimations. As shown, all these three datasets well reconstructed temporal variations of $PM_{2.5}$ from 2019 to 2020. Temporally, LGHAP and TAP are continuous while CHAP suffers from significant data gaps because no gap filling method was applied when generating the dataset. Compared with the other two datasets, LGHAP $PM_{2.5}$ data had a better agreement with ground-based $PM_{2.5}$ measurements. This high accuracy could be partially due to the fusion of *in situ* $PM_{2.5}$ data measured at adjacent sites via the OI method. Figure S5 compares $PM_{2.5}$ time series from LGHAP with $PM_{2.5}$ measurements sampled at five United States Embassy in China. It is indicative that historical $PM_{2.5}$ variations over these five cities were well reconstructed in LGHAP, even over years before 2014 at which $PM_{2.5}$ measurements from state-control monitoring sites were not available. Note $PM_{2.5}$ estimations appeared to significantly underestimate $PM_{2.5}$ concentration sampled at the Embassy in Beijing before 2013.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Considering the reconstructed AOD time series agreed well with AERONET AOD in Beijing (Figure 3a), and the model performed well in predicting historical $PM_{2.5}$ in Shanghai during the synchronous time period (Figure S5b), we are more willing to attribute this issue to significant $PM_{2.5}$ overestimations by the US Embassy during that period. Overall, these independent validation results collectively indicate a good accuracy of $PM_{2.5}$ in LGHAP dataset.

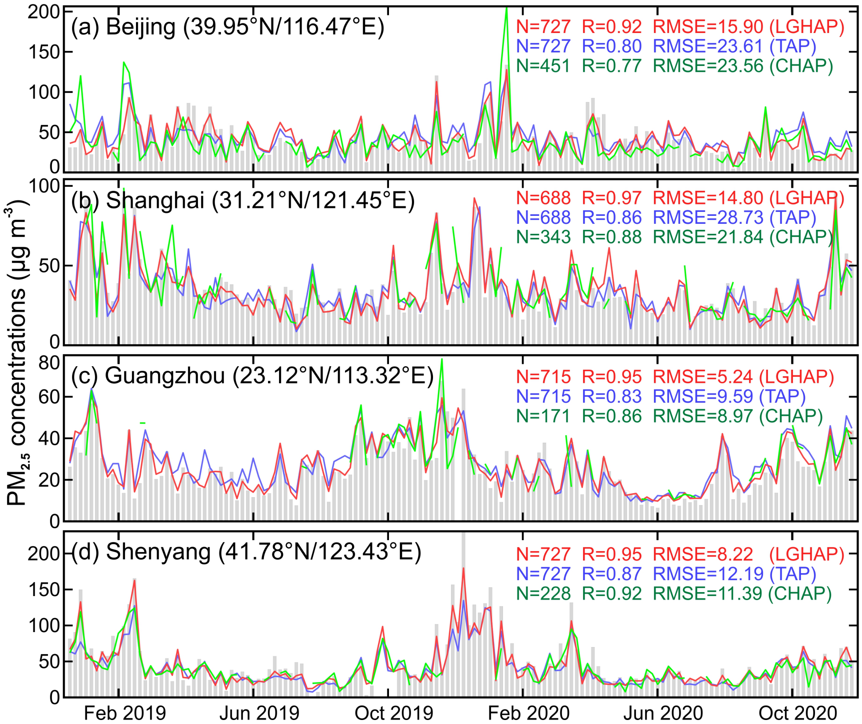



**Figure 7.** Comparison of $PM_{2.5}$ concentration time series between LGHAP (red line) and two open datasets (blue: TAP, green: CHAP). Here, hourly $PM_{2.5}$ concentrations measured by four United States Embassy in China from 2019 to 2020 (grey bar) were used as an independent $PM_{2.5}$ dataset to validate these three daily products. CHAP and TAP are two open access datasets providing PM2.5 concentration that were created by Wei et al. (2021) and Geng et al. (2021) respectively.

In Figure 8 we compared the spatial distribution of $PM_{2.5}$ that was reconstructed by different datasets. Compared to LGHAP and TAP, $PM_{2.5}$ data from CHAP are not gap free since the spatial coverage is determined by the AOD data coverage in the MAIAC product. Compared to TAP, LGHAP $PM_{2.5}$ data have a finer resolution (1 km versus 10 km), enabling us to examine $PM_{2.5}$ variations in

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


space with more details. Overall, LGHAP has a better performance in reconstructing $PM_{2.5}$ spatial distributions than the other two datasets. Reasons could be attributed to the following two aspects. Firstly, in situ $PM_{2.5}$ measurements were fused with gridded $PM_{2.5}$ estimations using the OI method when generating the final $PM_{2.5}$ product in LGHAP. This can help correct modeling biases in original $PM_{2.5}$ estimations. Secondly, a set of satellite-based AOD retrievals were incorporated when generating the full-coverage AOD product, which greatly helps reduce large biases in numerical AOD simulations, yielding more accurate $PM_{2.5}$ estimations in turn. This also highlights the great advantages of using big data analytics methods to advance air pollution assessment.

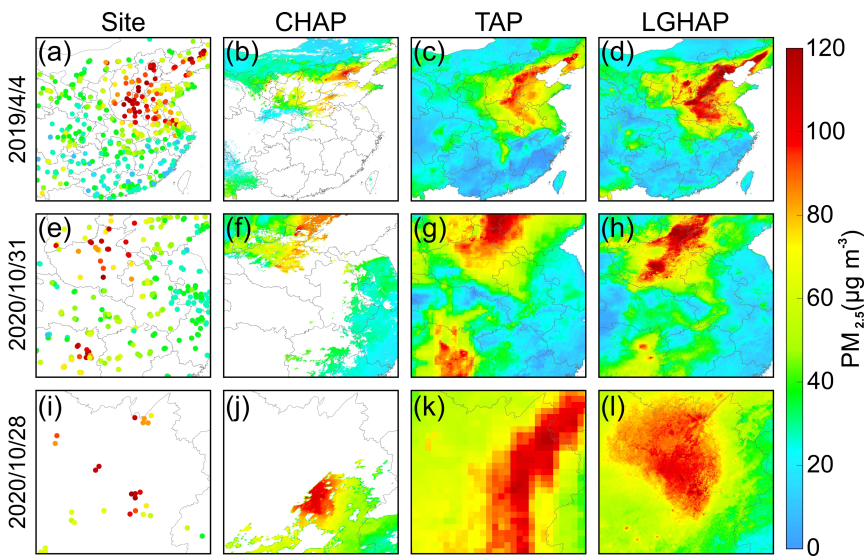



**Figure 8.** Comparison of $PM_{2.5}$ distribution reconstructed by different $PM_{2.5}$ concentration datasets. From the left to right, it shows in situ $PM_{2.5}$ concentration measurements, CHAP, TAP, and LGHAP, respectively.

To illustrate the fine resolution of LGHAP dataset, we compared the annual mean $PM_{10}$ concentration in 2019 with the proportion of impervious surface that was derived from 30-m resolution land cover dataset in eastern China. As shown in Figure 9, the finer resolution of LGHAP dataset enables us to easily recognize the "hot spot" regions with high $PM_{10}$ loading. By referring to the impervious surface distribution on the right, we found that these hot spots are mainly over cities and towns, indicative of the presence of pollution island in urban regions. Owing to the involvement of such high-resolution datasets, the spatial details of $PM_{2.5}$ and $PM_{10}$ can be then well recognized in

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021

Open Access Earth System Science Data Discussions

LGHAP. The finer spatial resolution advantage of the LGHAP dataset can be also demonstrated by comparisons of spatial distribution of annual mean $PM_{2.5}$ concentration that was revealed by four different datasets shown in Figure S6.

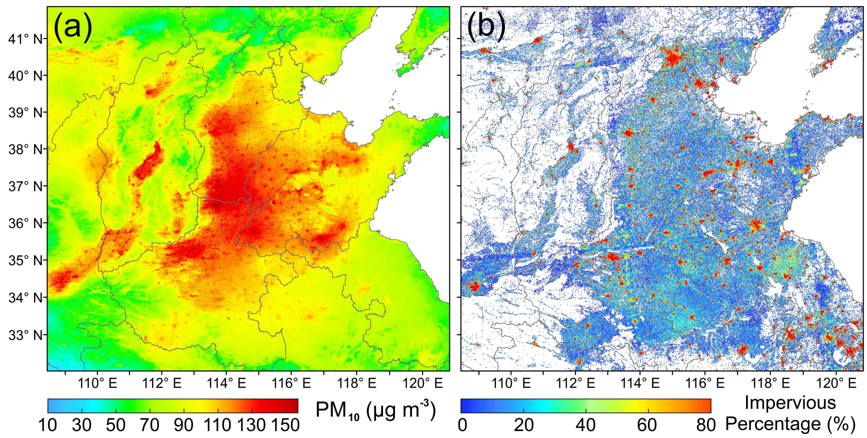



**Figure 9.** Comparison of annual mean $PM_{10}$ concentration with the proportion of areas coved by impervious surface in eastern China.

### 4.3 Long-term trends of haze pollution in China from 2000 to 2020

The aerosol pollution trends in China can be better examined by taking advantage of LGHAP dataset given long temporal coverage, gap free and high-resolution superiorities. Severe haze pollutions such as $PM_{2.5}$ are oftentimes observed during the wintertime (September to February). In this study, we first calculated wintertime mean $PM_{2.5}$ concentration in China from 2000 to 2020. As shown in Figure 10, severe wintertime haze pollution events were mainly observed in North China, especially over the adjacent region in Hebei-Shandong-Henan provinces. In addition, Sichuan basin and Fenwei plain also suffered from severe haze pollution during the wintertime. Temporally, severe haze pollution events occurred mainly from the late 2002 to early 2017, which were significantly reduced after 2017. Similar pattern can be also inferred from $PM_{10}$ concentration distributions shown in Figure S7.

Figure 11 shows the temporal variations of the proportion of land areas covered by $PM_{2.5}$ concentration exceeding 35 μg $m^{-3}$ (the national ambient air quality standard for 24-hour $PM_{2.5}$ concentration given in GB 3095-2012). As shown in Figure 11a, severe $PM_{2.5}$ pollution occurred mainly during the wintertime in China, as more than one-third land areas (indicated by the blue lines)

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



were exposed to hazardous $PM_{2.5}$ pollutants. Meanwhile, an apparent inflection was observed in 2007, after which the number of episode days decreased drastically at more than one-third land area covered by $PM_{2.5}$ concentration exceeding 35 μg $m^{-3}$. According to the proportion of land area covered with annual mean $PM_{2.5}$ concentration greater than 35 μg $m^{-3}$, the variation of haze pollution in China can be generally divided into three different periods during the past two-decades (Figure 11b). As indicated, an increasing trend was observed from 2000 to 2007, during which land areas covered by $PM_{2.5}$ concentration greater than 35 μg $m^{-3}$ had increased to near 40% at a pace of 1.04% $a^{-1}$. The second period was from 2008 to 2013, during which the land area coverage ratio decreased at a rate of –0.21% $a^{-1}$. The third period started from 2014, after which the land area covered with $PM_{2.5}$ concentration more than 35 μg $m^{-3}$ had decreased drastically, at a rate of –2.23% $a^{-1}$.

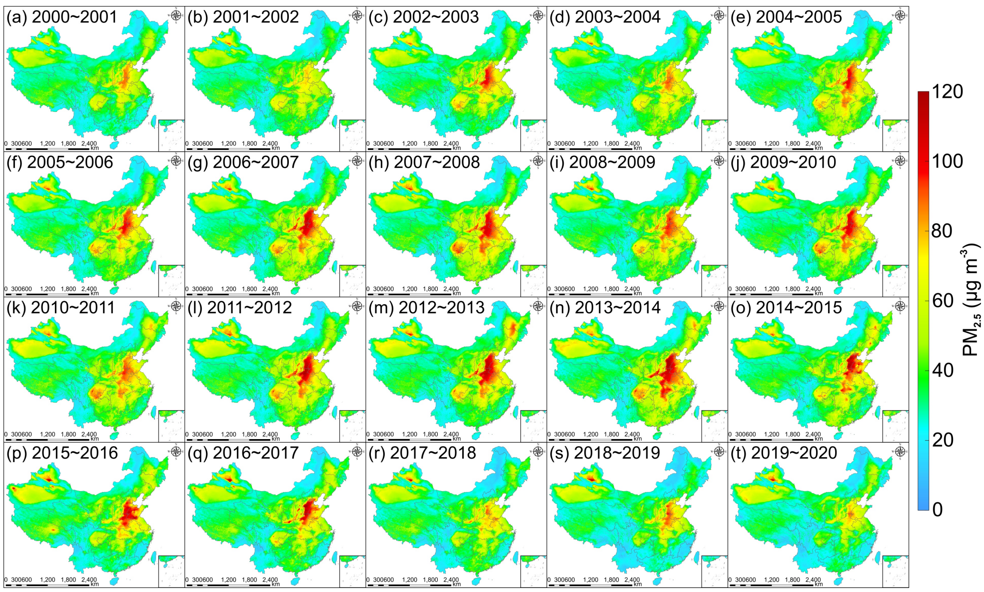



**Figure 10.** Spatial distribution of wintertime (September to February) averaged $PM_{2.5}$ concentration from LGHAP during 2000 to 2020 in China.

Figure 11c–e presents the linear trend of $PM_{2.5}$ concentration during these three specific periods, from which we observed that significant $PM_{2.5}$ variations occurred mainly over eastern parts of the country where resides two-thirds of the population. A near ubiquitous $PM_{2.5}$ increasing trend was observed during 2000–2007, with significant increase (>1.0 μg $m^{-3}$ $a^{-1}$) mainly observed in eastern China. During the second period, $PM_{2.5}$ concentration over most regions shows a small decreasing

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




trend except in the Ji-Lu-Yu region where an increasing trend was still observed. Apparent decreasing trend was observed over most parts of the country after 2014, indicative of significant reductions in $PM_{2.5}$ loading across China. This trend distribution is in line with our previous study that was derived using the annual mean $PM_{2.5}$ concentration dataset generated by the Dalhousie University (Bai et al., 2019b). However, differences were still observed in terms of the regions where significant decreasing trends were present. Most significant decreasing trends were mainly observed in Sichuan basin and Pearl River Delta in the previous study. However, regions with drastic $PM_{2.5}$ decrease were found mainly in the North China where severe haze pollution events were oftentimes reported. Similar variation patterns can be also inferred from $PM_{10}$ (Figure S8) and AOD (Figure S9). Overall, the LGHAP dataset provides us a gridded perspective to better examine long-term variations of haze pollution in China during the past two decades.

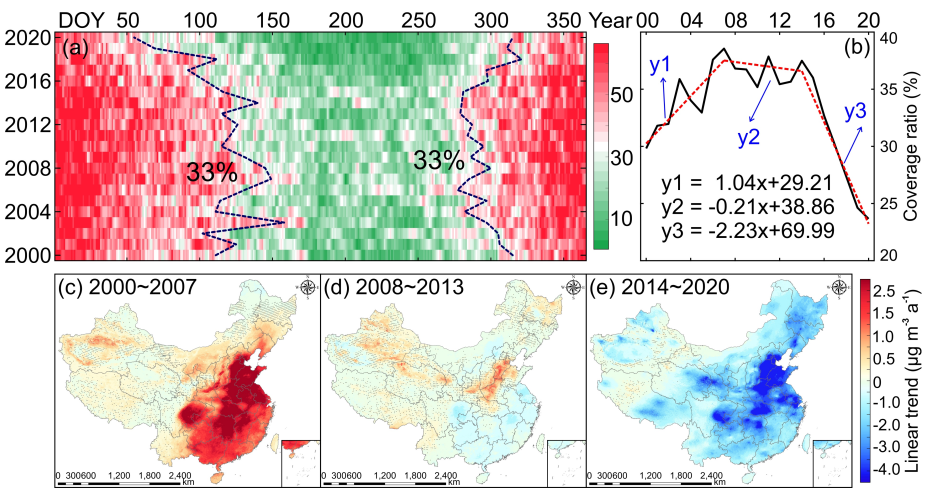



**Figure 11.** Temporal variations of the proportion of land areas covered with $PM_{2.5}$ concentration exceeding 35 μg $m^{-3}$ and $PM_{2.5}$ trends during three different periods. (a) Temporal variations of the land coverage ratio with daily $PM_{2.5}$ concentration exceeding 35 μg $m^{-3}$ from 2000 to 2000. (b) same as (a) but for annual mean $PM_{2.5}$ concentration. (c–e) $PM_{2.5}$ trends during periods of 2000–2007, 2008–2013, and 2014–2020. The dotted regions imply trend estimations are statistically insignificant at the 95% confidence interval.

### 4.4 Population exposure to $PM_{2.5}$ pollution in China

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Earth System Science Data
Open Access
Discussions

By taking advantage of fine resolution LGHAP $PM_{2.5}$ concentration and gridded population data, population exposure to $PM_{2.5}$ pollution across China over the past two decades were estimated. Figure 12 shows the spatial distribution of population weighted $PM_{2.5}$ concentration and the proportion of population exposed to $PM_{2.5}$ concentration greater than 35 μg $m^{-3}$. As shown, spatial distribution of population weighted $PM_{2.5}$ concentration resembles the spatial pattern of annual mean $PM_{2.5}$ concentration, with high values observed mainly in eastern and central China as well as northwest China. Nonetheless, $PM_{2.5}$ sources in these two areas could be different. In northwest China, natural emissions could be the dominant source since very limited population resides there. In contrast, most population lives in eastern and central China with highly developed economy, and anthropogenic emissions thus might play more important roles in $PM_{2.5}$ formation (Xin et al., 2015; Yang et al., 2011). In regard to the proportion of population exposed to the ambient with $PM_{2.5}$ concentration greater than 35 μg $m^{-3}$, we observed that the annual mean population ratio exposure to unhealthy $PM_{2.5}$ increased gradually from 50.60% in 2000 to 65.72% in 2007. During 2007–2014, the ratio varied with small changes (<5%), whereas a drastic decline was observed after 2014, with the annual mean proportion of population exposed to unhealthy $PM_{2.5}$ was reduced from 63.81% in 2014 to 34.03% in 2020, even though the total population was increased from 1.37 billion to 1.41 billion during the synchronous period. Nonetheless, more than one-third population was still exposed to unhealthy $PM_{2.5}$, highlighting the requirement of further emission reduction actions to manage haze pollutions in China.

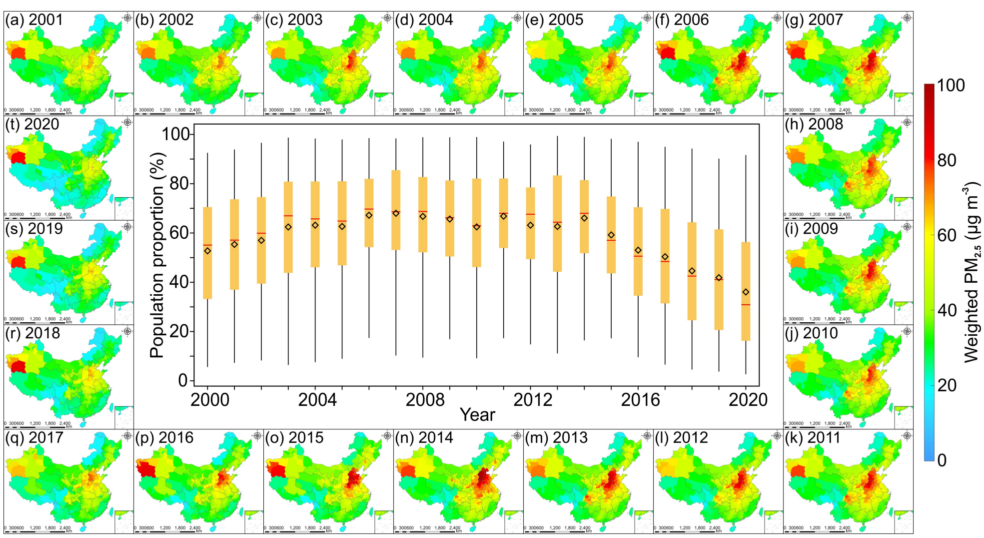


**Figure 12.** Spatial distribution of population weighted $PM_{2.5}$ concentration and the proportion of population exposed to $PM_{2.5}$ concentration greater than 35 μg $m^{-3}$. Annual and daily LGHAP $PM_{2.5}$ concentration data were used for the calculation of weighted $PM_{2.5}$ and the proportion of population exposure, respectively. The diamond and red line indicate the annual mean and median population proportion, respectively.

## 5 Data availability

The LGHAP dataset, consisting of gap free AOD, $PM_{2.5}$, and $PM_{10}$ concentration with daily 1-km resolution from 2000 to 2020, are all publicly accessible. The daily map was provided in the NetCDF format and data in each individual year were archived in a zip file. For AOD, the dataset has a disk storage size of near 27 GB in total, which can be found at https://doi.org/10.5281/zenodo.5652257 (Bai et al., 2021a). $PM_{2.5}$ (38 GB) and $PM_{10}$ (48 GB) concentration data can be acquired from https://doi.org/10.5281/zenodo.5652265 (Bai et al., 2021b) and https://doi.org/10.5281/zenodo.5652263 (Bai et al., 2021c), respectively. Additionally, monthly and annual mean datasets were also provided, which can be acquired from https://doi.org/10.5281/zenodo.5655797 (Bai et al., 2021d) and https://doi.org/10.5281/zenodo.5655807 (Bai et al., 2021e), respectively.

## 6 Conclusion

In this study, a big data analytics method was developed for generating a LGHAP dataset to advance research in earth system science and environment management. With integrative efforts of fusing AOD features extracted from a set of AOD data tensors and knowledge transfer in statistical data mining from diverse air quality indicators, a LGHAP aerosol dataset providing 21-year-long (2000–2020) gap-free AOD, $PM_{2.5}$, and $PM_{10}$ concentration data with daily 1-km resolution in China, was generated. Gap-filled AOD imageries were firstly generated by reconstructing AOD distribution in $AOD_{Terra}$ via fusing AOD features recognized from diversified satellites and numerical models as well as in situ data through tensor completion. Compared to ground-based AOD measurements, the gap-filled AOD data exhibit a satisfying prediction accuracy and good performance in delineating AOD variations over space and time. To our knowledge, this is the first thrust of generating long-term high-resolution AOD dataset with gap free nature in China.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




$PM_{2.5}$ and $PM_{10}$ concentration data were then estimated using an ensemble learning approach by taking advantage of the generated gap-free AOD imageries. Ground validation results also indicate good accuracies of these two gridded products, showing a comparable bias level with many previous studies. Compared with other open access daily $PM_{2.5}$ concentration datasets, the LGHAP $PM_{2.5}$ dataset performs well due to the vantage of having gap free and fine resolution products. With this gap free and high-resolution dataset, the long-term variation trend of haze pollution in China over the past two decades was examined, and apparent inflections were observed in 2007 and 2014, at which $PM_{2.5}$ concentration was found to turn from an increasing path to decreasing in 2007 with a more drastic decline observed starting from 2014. Moreover, the LGHAP dataset provides us a gridded perspective to assess two-decade long population exposure to $PM_{2.5}$ pollution in China. In spite of a drastic decline in population exposure, there are still more than one-third population exposed to unhealthy $PM_{2.5}$ pollutants, highlighting the requirement of long-lasting actions to continue $PM_{2.5}$ related emission reduction.

Overall, these three gridded LGHAP aerosol products provide a long-term perspective on aerosol changes over different regions of China, and users are encouraged to use the LGHAP dataset to assess aerosol impacts on public health, air quality, climate, and ecosystem. The dataset has been publicly released online and is freely accessible via the links provided above. In addition to the LGHAP dataset, Python, Matlab, R, and IDL codes that can be used to read and visualize these data were provided as well. Global scale dataset is on the track and will be released to the public soon.

**Author contributions**

The study was completed with cooperation between all authors. KB, KL, JG, ZL and N.B.C conceived of the idea behind generating the LGHAP dataset. KL, KB, and ZT developed the method and KB wrote the paper. KL, KB, K.T.L, and MM conducted the data analyses. JG and ZL provided atmospheric visibility and in situ AOD data, respectively. All authors discussed the results and proofread the paper.

**Competing interests**

The authors declare that they have no conflict of interest.

**Acknowledgments**

The authors are grateful to all organizations and groups for providing essential datasets that were used in this study. The MAIAC AOD was acquired from https://lpdaac.usgs.gov/products/mcd19a2v006/. The MISR AOD was acquired from https://asdc.larc.nasa.gov/project/MISR. The VIIRS AOD was acquired from https://earthdata.nasa.gov/earth-observation-data/near-real-time/download-nrt-data/viirs-nrt. The AATSR AOD was acquired from https://climate.esa.int/en/projects/aerosol/data/. The POLDER AOD was acquired from https://www.grasp-open.com/products/polder-data-release/. The aerosol diagnostics including AOD and aerosol components from MERRA-2 were acquired from https://disc.gsfc.nasa.gov/datasets/M2T1NXAER_5.12.4/summary?keywords=MERRA2. AOD from AERONET was acquired from https://aeronet.gsfc.nasa.gov/new_web/aerosols.html. Meteorological factors were retrieved from the latest ERA-5 reanalysis and can be reached at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview. Atmospheric visibility data were acquired from the national meteorological information center at http://data.cma.cn/en. Ground-based air pollutants concentration was acquired from https://air.cnemc.cn:18007/. Gridded Population data were acquired from https://www.worldpop.org/ while DEM was acquired from https://www.resdc.cn/. Monthly NDVI data were acquired from https://lpdaac.usgs.gov/products/mod13a3v061/. Land cover data were acquired from http://www.globallandcover.com/defaults.html?src=/Scripts/map/defaults/browse.html&head=browse&type=data and https://zenodo.org/record/4417810#.YSxD844zYuW.

**Financial support**

This study was supported by the National Natural Science Foundation of China (grants 42171309 and 41701413), and the Shanghai Committee of Science and Technology (grant 20ZR1415900).

## References

Bai, K., Chang, N.-B. and Chen, C.-F.: Spectral Information Adaptation and Synthesis Scheme for Merging Cross-Mission Ocean Color Reflectance Observations From MODIS and VIIRS, IEEE Trans. Geosci. Remote Sens., 54(1), 311–329, doi:10.1109/TGRS.2015.2456906, 2016.

Bai, K., Li, K., Chang, N.-B. and Gao, W.: Advancing the prediction accuracy of satellite-based PM2.5 concentration mapping: A perspective of data mining through in situ PM2.5 measurements, Environ. Pollut., 254, 113047, doi:10.1016/j.envpol.2019.113047, 2019a.

Bai, K., Ma, M., Chang, N.-B. and Gao, W.: Spatiotemporal trend analysis for fine particulate matter concentrations in China using high-resolution satellite-derived and ground-measured PM2.5 data, J. Environ. Manage., 233, 530–542, doi:10.1016/j.jenvman.2018.12.071, 2019b.

Bai, K., Li, K., Wu, C., Chang, N.-B. and Guo, J.: A homogenized daily in situ PM2.5 concentration dataset from the national air quality monitoring network in China, Earth Syst. Sci. Data, 12(4), 3067–3080, doi:10.5194/essd-12-3067-2020, 2020a.

Bai, K., Li, K., Guo, J., Yang, Y. and Chang, N.-B.: Filling the gaps of in situ hourly PM2.5 concentration data with the aid of empirical orthogonal function analysis constrained by diurnal cycles, Atmos. Meas. Tech., 13(3), 1213–1226, doi:10.5194/amt-13-1213-2020, 2020b.

Bai, K., Li, K., Guo, J. and Chang, N.-B.: Multiscale and multisource data fusion for full-coverage PM2.5 concentration mapping: Can spatial pattern recognition come with modeling accuracy?, ISPRS J. Photogramm. Remote Sens., 2021. in revision

Bai, K., Li, K. Tan, Z., Han, D., and Guo, J.: Daily 1-km gap-free AOD grids in China, v1 (2000–2020) [data set], https://doi.org/10.5281/zenodo.5652257, 2021a.

Bai, K., Li, K. Tan, Z., Han, D., and Guo, J.: Daily 1-km gap-free $PM_{2.5}$ grids in China, v1 (2000–2020) [data set], https://doi.org/10.5281/zenodo.5652265, 2021b.

Bai, K., Li, K. Tan, Z., Han, D., and Guo, J.: Daily 1-km gap-free $PM_{10}$ grids in China, v1 (2000–2020) [data set], https://doi.org/10.5281/zenodo.5652263, 2021c.

Bai, K., Li, K. Tan, Z., Han, D., and Guo, J.: Monthly averaged 1-km gap-free AOD, $PM_{2.5}$ and $PM_{10}$ grids in China, v1 (2000–2020) [data set], https://doi.org/10.5281/zenodo.5655797, 2021d.

Bai, K., Li, K. Tan, Z., Han, D., and Guo, J.: Annual mean 1-km gap-free AOD, $PM_{2.5}$ and $PM_{10}$ grids in China, v1 (2000–2020) [data set], https://doi.org/10.5281/zenodo.5655807, 2021e.

Beckers, J. M. and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Datasets, J. Atmos. Ocean. Technol., 20(12), 1839–1856, doi:10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2, 2003.

Bi, J., Belle, J. H., Wang, Y., Lyapustin, A. I., Wildani, A. and Liu, Y.: Impacts of snow and cloud covers on satellite-derived PM2.5 levels, Remote Sens. Environ., 221(October), 665–674, doi:10.1016/j.rse.2018.12.002, 2018.

Chang, N.-B., Bai, K. and Chen, C.-F.: Smart Information Reconstruction via Time-Space-Spectrum Continuum for Cloud Removal in Satellite Images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 8(5), 1898–1912, doi:10.1109/JSTARS.2015.2400636, 2015.

Che, H., Yang, L., Liu, C., Xia, X., Wang, Y., Wang, H., Wang, H., Lu, X. and Zhang, X.: Long-term validation of MODIS C6 and C6.1 Dark Target aerosol products over China using CARSNET and AERONET, Chemosphere, 236, 124268, doi:10.1016/j.chemosphere.2019.06.238, 2019.

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J. and Guo, Y.: A machine learning method to estimate PM 2.5 concentrations across China with remote sensing, meteorological and land use information, Sci. Total Environ., 636, 52–60, doi:10.1016/j.scitotenv.2018.04.251, 2018.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J. and Schwartz, J.: An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution, Environ. Int., 130, 104909, doi:10.1016/j.envint.2019.104909, 2019.

van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C. and Villeneuve, P. J.: Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application, Environ. Health Perspect., 118(6), 847–855, doi:10.1289/ehp.0901623, 2010.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M. and Winker, D. M.: Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors, Environ. Sci. Technol., 50(7), 3762–3772, doi:10.1021/acs.est.5b05833, 2016.

Fang, X., Zou, B., Liu, X., Sternberg, T. and Zhai, L.: Satellite-based ground PM2.5 estimation using timely structure adaptive modeling, Remote Sens. Environ., 186, 152–163, doi:10.1016/j.rse.2016.08.027, 2016.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021



Folch-Fortuny, A., Arteaga, F. and Ferrer, A.: PCA model building with missing data: New proposals and a comparative study, Chemom. Intell. Lab. Syst., 146, 77–88, doi:10.1016/j.chemolab.2015.05.006, 2015.

Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., Fowler, D., Koren, I., Langford, B., Lohmann, U., Nemitz, E., Pandis, S., Riipinen, I., Rudich, Y., Schaap, M., Slowik, J. G., Spracklen, D. V., Vignati, E., Wild, M., Williams, M. and Gilardoni, S.: Particulate matter, air quality and climate: lessons learned and future needs, Atmos. Chem. Phys., 15(14), 8217–8299, doi:10.5194/acp-15-8217-2015, 2015.

Gao, M., Beig, G., Song, S., Zhang, H., Hu, J., Ying, Q., Liang, F., Liu, Y., Wang, H., Lu, X., Zhu, T., Carmichael, G. R., Nielsen, C. P. and McElroy, M. B.: The impact of power generation emissions on ambient PM2.5 pollution and human health in China and India, Environ. Int., 121(August), 250–259, doi:10.1016/j.envint.2018.09.015, 2018.

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., Peng, Y., Huang, X., He, K. and Zhang, Q.: Tracking Air Pollution in China: Near Real-Time PM 2.5 Retrievals from Multisource Data Fusion, Environ. Sci. Technol., 55(17), 12106–12115, doi:10.1021/acs.est.1c01863, 2021a.

Geng, G., Xiao, Q., Liu, S., Liu, X., Cheng, J., Zheng, Y., Xue, T., Tong, D., Zheng, B., Peng, Y., Huang, X., He, K. and Zhang, Q.: Tracking Air Pollution in China: Near Real-Time PM 2.5 Retrievals from Multisource Data Fusion, Environ. Sci. Technol., acs.est.1c01863, doi:10.1021/acs.est.1c01863, 2021b.

Goldberg, D. L., Gupta, P., Wang, K., Jena, C., Zhang, Y., Lu, Z. and Streets, D. G.: 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(November 2018), 443–452, doi:10.1016/j.atmosenv.2018.11.049, 2019.

Guo, J., Su, T., Li, Z., Miao, Y., Li, J., Liu, H., Xu, H., Cribb, M. and Zhai, P.: 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, doi:10.1002/2017GL073533, 2017.

He, Q., Gu, Y. and Zhang, M.: Spatiotemporal trends of PM2.5 concentrations in central China from 2003 to 2018 based on MAIAC-derived high-resolution data, Environ. Int., 137(August 2019), 105536, doi:10.1016/j.envint.2020.105536, 2020.

Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., Estes, M. G., Estes, S. M., Quattrochi, D. A., Puttaswamy, S. J. and Liu, Y.: Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model, Remote Sens. Environ., 140, 220–232, doi:10.1016/j.rse.2013.08.032, 2014.

Huang, C., Hu, J., Xue, T., Xu, H. and Wang, M.: High-Resolution Spatiotemporal Modeling for Ambient PM2.5Exposure Assessment in China from 2013 to 2019, Environ. Sci. Technol., 55(3), 2152–2162, doi:10.1021/acs.est.0c05815, 2021.

Jun, C., Ban, Y. and Li, S.: Open access to Earth land-cover map, Nature, 514(7523), 434–434, doi:10.1038/514434c, 2014.

Kolda, T. G. and Bader, B. W.: Tensor Decompositions and Applications, SIAM Rev., 51(3), 455–500, doi:10.1137/07070111X, 2009.

de Leeuw, G., Sogacheva, L., Rodriguez, E., Kourtidis, K., Georgoulias, A. K., Alexandri, G., Amiridis, V., Proestakis, E., Marinou, E., Xue, Y. and van der A, R.: Two decades of satellite observations of AOD over mainland China using ATSR-2, AATSR and MODIS/Terra: data set evaluation and large-scale patterns, Atmos. Chem. Phys., 18(3), 1573–1592, doi:10.5194/acp-18-1573-2018, 2018.

Li, J., Li, C. and Zhao, C.: Different trends in extreme and median surface aerosol extinction coefficients over China inferred from quality-controlled visibility data, Atmos. Chem. Phys., 18(5), 3289–3298, doi:10.5194/acp-18-3289-2018, 2018a.

Li, L., Zhang, J., Meng, X., Fang, Y., Ge, Y., Wang, J., Wang, C., Wu, J. and Kan, H.: Estimation of PM2.5 concentrations at a high spatiotemporal resolution using constrained mixed-effect bagging models with MAIAC aerosol optical depth, Remote Sens. Environ., 217(January), 573–586, doi:10.1016/j.rse.2018.09.001, 2018b.

Li, L., Franklin, M., Girguis, M., Lurmann, F., Wu, J., Pavlovic, N., Breton, C., Gilliland, F. and Habre, R.: Spatiotemporal imputation of MAIAC AOD using deep learning with downscaling, Remote Sens. Environ., 237(October 2019), 111584, doi:10.1016/j.rse.2019.111584, 2020.

Li, K., Bai, K., Li, Z., Guo, J. and Chang, N.-B.: Synergistic Data Fusion of Multimodal AOD and Air Quality Data for Near Real-Time Full Coverage Air Pollution Assessment, J. Environ. Manage., 2021. In revision

Li, Z., Zhang, Y., Shao, J., Li, B., Hong, J., Liu, D., Li, D., Wei, P., Li, W., Li, L., Zhang, F., Guo,

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


J., Deng, Q., Wang, B., Cui, C., Zhang, W., Wang, Z., Lv, Y., Xu, H., Chen, X., Li, L. and Qie, L.: Remote sensing of atmospheric particulate mass of dry PM2.5 near the ground: Method validation using ground-based measurements, Remote Sens. Environ., 173, 59–68, doi:10.1016/j.rse.2015.11.019, 2016.

Li, Z., Wang, Y., Guo, J., Zhao, C., Cribb, M. C., Dong, X., Fan, J., Gong, D., Huang, J., Jiang, M., Jiang, Y., Lee, S. S., Li, H., Li, J., Liu, J., Qian, Y., Rosenfeld, D., Shan, S., Sun, Y., Wang, H., Xin, J., Yan, X., Yang, X., Yang, X. qun, Zhang, F. and Zheng, Y.: East Asian Study of Tropospheric Aerosols and their Impact on Regional Clouds, Precipitation, and Climate (EAST-AIRCPC), J. Geophys. Res. Atmos., 124(23), 13026–13054, doi:10.1029/2019JD030758, 2019.

Lin, C., Li, Y., Lau, A. K. H., Deng, X., Tse, T. K. T., Fung, J. C. H., Li, C., Li, Z., Lu, X., Zhang, X. and Yu, Q.: 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, doi:10.1016/j.rse.2016.03.023, 2016.

Liu, M., Bi, J. and Ma, Z.: Visibility-Based PM 2.5 Concentrations in China: 1957–1964 and 1973–2014, Environ. Sci. Technol., 51(22), 13161–13169, doi:10.1021/acs.est.7b03468, 2017.

Liu, Y., Paciorek, C. J. and Koutrakis, P.: Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information, Environ. Health Perspect., 117(6), 886–892, doi:10.1289/ehp.0800123, 2009.

Lyapustin, A., Martonchik, J., Wang, Y., Laszlo, I. and Korkin, S.: Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables, J. Geophys. Res. Atmos., 116(3), doi:10.1029/2010JD014985, 2011.

Lyapustin, A., Wang, Y., Korkin, S. and Huang, D.: MODIS Collection 6 MAIAC algorithm, Atmos. Meas. Tech., 11(10), 5741–5765, doi:10.5194/amt-11-5741-2018, 2018.

Lyu, B., Hu, Y., Zhang, W., Du, Y., Luo, B., Sun, X., Sun, Z., Deng, Z., Wang, X., Liu, J., Wang, X. and Russell, A. G.: Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: Application in China to Estimate Spatiotemporally-Resolved PM2.5 Exposure Fields in 2014-2017, Environ. Sci. Technol., 53(13), 7306–7315, doi:10.1021/acs.est.9b01117, 2019.

Ma, Z., Hu, X., Huang, L., Bi, J. and Liu, Y.: Estimating Ground-Level PM 2.5 in China Using Satellite Remote Sensing, Environ. Sci. Technol., 48(13), 7436–7444, doi:10.1021/es5009399, 2014.

Ma, Z., Hu, X., Sayer, A. M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L. and Liu,

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Open Access Earth System Science Data Discussions

Y.: Satellite-based spatiotemporal trends in PM2.5 concentrations: China, 2004-2013, Environ. Health Perspect., 124(2), 184–192, doi:10.1289/ehp.1409481, 2016.

Park, S., Lee, J., Im, J., Song, C. K., Choi, M., Kim, J., Lee, S., Park, R., Kim, S. M., Yoon, J., Lee, D. W. and Quackenbush, L. J.: Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models, Sci. Total Environ., 713, 136516, doi:10.1016/j.scitotenv.2020.136516, 2020.

Shen, F., Zhang, L., Jiang, L., Tang, M., Gai, X., Chen, M. and Ge, X.: Temporal variations of six ambient criteria air pollutants from 2015 to 2018, their spatial distributions, health risks and relationships with socioeconomic factors during 2018 in China, Environ. Int., 137(February), 105556, doi:10.1016/j.envint.2020.105556, 2020.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E. and Faloutsos, C.: Tensor Decomposition for Signal Processing and Machine Learning, IEEE Trans. Signal Process., 65(13), 3551–3582, doi:10.1109/TSP.2017.2690524, 2017.

Sogacheva, L., Popp, T., Sayer, A. M., Dubovik, O., Garay, M. J., Heckel, A., Christina Hsu, N., Jethva, H., Kahn, R. A., Kolmonen, P., Kosmale, M., De Leeuw, G., Levy, R. C., Litvinov, P., Lyapustin, A., North, P., Torres, O. and Arola, A.: Merging regional and global aerosol optical depth records from major available satellite products, Atmos. Chem. Phys., 20(4), 2031–2056, doi:10.5194/acp-20-2031-2020, 2020.

Sun, J.-L., Jing, X., Chang, W.-J., Chen, Z.-X. and Zeng, H.: Cumulative health risk assessment of halogenated and parent polycyclic aromatic hydrocarbons associated with particulate matters in urban air, Ecotoxicol. Environ. Saf., 113, 31–37, doi:10.1016/j.ecoenv.2014.11.024, 2015.

Sun, Z., Chang, N. Bin, Chen, C. F., Mostafiz, C. and Gao, W.: Ensemble learning via higher order singular value decomposition for integrating data and classifier fusion in water quality monitoring, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 14, 3345–3360, doi:10.1109/JSTARS.2021.3055798, 2021.

Tang, Q., Bo, Y. and Zhu, Y.: Spatiotemporal fusion of multiple-satellite aerosol optical depth (AOD) products using Bayesian maximum entropy method, J. Geophys. Res. Atmos., 121(8), 4034–4048, doi:10.1002/2015JD024571, 2016.

Tucker, L. R.: Some mathematical notes on three-mode factor analysis, Psychometrika, 31(3), 279–311, doi:10.1007/BF02289464, 1966.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Wang, B., Yuan, Q., Yang, Q., Zhu, L., Li, T. and Zhang, L.: Estimate hourly PM2.5 concentrations from Himawari-8 TOA reflectance directly using geo-intelligent long short-term memory network, Environ. Pollut., 271, 116327, doi:10.1016/j.envpol.2020.116327, 2021a.

Wang, Q., Shen, Y., and Zhang, J. Q.: A nonlinear correlation measure for multivariable data set, Phys. D, 3–4, 287–295, doi:10.1016/j.physd.2004.11.001, 2005.

Wang, Y., Yuan, Q., Li, T., Shen, H., Zheng, L. and Zhang, L.: Large-scale MODIS AOD products recovery: Spatial-temporal hybrid fusion considering aerosol variation mitigation, ISPRS J. Photogramm. Remote Sens., 157(July), 1–12, doi:10.1016/j.isprsjprs.2019.08.017, 2019.

Wang, Y., Yuan, Q., Li, T., Tan, S. and Zhang, L.: Full-coverage spatiotemporal mapping of ambient PM2.5 and PM10 over China from Sentinel-5P and assimilated datasets: Considering the precursors and chemical compositions, Sci. Total Environ., 793, 148535, doi:10.1016/j.scitotenv.2021.148535, 2021b.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L. and Cribb, M.: Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach, Remote Sens. Environ., 231(May), 111221, doi:10.1016/j.rse.2019.111221, 2019a.

Wei, J., Li, Z., Peng, Y. and Sun, L.: MODIS Collection 6.1 aerosol optical depth products over land and ocean: validation and comparison, Atmos. Environ., 201, 428–440, doi:10.1016/j.atmosenv.2018.12.004, 2019b.

Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H. and Song, Y.: Improved 1 km resolution PM2.5 estimates across China using enhanced space – time extremely randomized trees, Atmos. Chem. Phys., 20, 3273–3289, 2020a.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T. and Cribb, M.: 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(January 2020), 112136, doi:10.1016/j.rse.2020.112136, 2021a.

Wei, X., Chang, N., Bai, K. and Gao, W.: Satellite remote sensing of aerosol optical depth: advances, challenges, and perspectives, Crit. Rev. Environ. Sci. Technol., 50(16), 1640–1725, doi:10.1080/10643389.2019.1665944, 2020b.

Wei, X., Bai, K., Chang, N. and Gao, W.: Multi-source hierarchical data fusion for high-resolution AOD mapping in a forest fire event, Int. J. Appl. Earth Obs. Geoinf., 102(May), 102366,

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021




doi:10.1016/j.jag.2021.102366, 2021b.

Xiao, Q., Zhang, H., Choi, M., Li, S., Kondragunta, S., Kim, J., Holben, B., Levy, R. C. and Liu, Y.: Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia, Atmos. Chem. Phys., 16(3), 1255–1269, doi:10.5194/acp-16-1255-2016, 2016.

Xiao, Q., Wang, Y., Chang, H. H., Meng, X., Geng, G., Lyapustin, A. and Liu, Y.: Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China, Remote Sens. Environ., 199(May), 437–446, doi:10.1016/j.rse.2017.07.023, 2017a.

Xiao, Q., Wang, Y., Chang, H. H., Meng, X., Geng, G., Lyapustin, A. and Liu, Y.: Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China, Remote Sens. Environ., 199, 437–446, doi:10.1016/j.rse.2017.07.023, 2017b.

Xiao, Q., Chang, H. H., Geng, G. and Liu, Y.: An Ensemble Machine-Learning Model to Predict Historical PM2.5 Concentrations in China from Satellite Data, Environ. Sci. Technol., doi:10.1021/acs.est.8b02917, 2018.

Xiao, Q., Geng, G., Liang, F., Wang, X., Lv, Z., Lei, Y., Huang, X., Zhang, Q., Liu, Y. and He, K.: Changes in spatial patterns of PM2. 5 pollution in China 2000 – 2018 : Impact of clean air policies, Environ. Int., 141(April), 105776, doi:10.1016/j.envint.2020.105776, 2020.

Xin, J., Wang, Y., Pan, Y., Ji, D., Liu, Z., Wen, T., Wang, Y., Li, X., Sun, Y., Sun, J., Wang, P., Wang, G., Wang, X., Cong, Z., Song, T., Hu, B., Wang, L., Tang, G., Gao, W., Guo, Y., Miao, H., Tian, S. and Wang, L.: The Campaign on Atmospheric Aerosol Research Network of China: CARE-China, Bull. Am. Meteorol. Soc., 96(7), 1137–1155, doi:10.1175/BAMS-D-14-00039.1, 2015.

Xu, H., Guang, J., Xue, Y., de Leeuw, G., Che, Y. H., Guo, J., He, X. W. and Wang, T. K.: A consistent aerosol optical depth (AOD) dataset over mainland China by integration of several AOD products, Atmos. Environ., 114, 48–56, doi:10.1016/j.atmosenv.2015.05.023, 2015.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T. and Zhang, Q.: 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(December 2018), 345–357, doi:10.1016/j.envint.2018.11.075, 2019.

Yang, F., Tan, J., Zhao, Q., Du, Z., He, K., Ma, Y., Duan, F., Chen, G. and Zhao, Q.: Characteristics of PM2.5 speciation in representative megacities and across China, Atmos. Chem.

https://doi.org/10.5194/essd-2021-404
Preprint. Discussion started: 18 November 2021


Phys., 11(11), 5207–5219, doi:10.5194/acp-11-5207-2011, 2011.

Yang, J. and Huang, X.: 30 m annual land cover and its dynamics in China from 1990 to 2019, Earth Syst. Sci. Data Discuss., 2021(April), 1–29, doi:https://doi.org/10.5194/essd-2021-7, 2021.

Yi-Lei Chen, Chiou-Ting Hsu and Liao, H.-Y. M.: Simultaneous Tensor Decomposition and Completion Using Factor Priors, IEEE Trans. Pattern Anal. Mach. Intell., 36(3), 577–591, doi:10.1109/TPAMI.2013.164, 2014.

Zhang, T., Zeng, C., Gong, W., Wang, L., Sun, K., Shen, H., Zhu, Z. and Zhu, Z.: Improving spatial coverage for Aqua MODIS AOD using NDVI-based multi-temporal regression analysis, Remote Sens., 9(4), doi:10.3390/rs9040340, 2017.

Zhang, Y., Gao, L., Cao, L., Yan, Z. and Wu, Y.: Decreasing atmospheric visibility associated with weakening winds from 1980 to 2017 over China, Atmos. Environ., 224(July 2019), 117314, doi:10.1016/j.atmosenv.2020.117314, 2020.

Zhao, C., Yang, Y., Fan, H., Huang, J., Fu, Y., Zhang, X., Kang, S., Cong, Z., Letu, H. and Menenti, M.: Aerosol characteristics and impacts on weather and climate over the Tibetan Plateau, Natl. Sci. Rev., 7(3), 492–495, doi:10.1093/nsr/nwz184, 2020.