

Tracking Air Pollution in China: Near Real-Time PM_{2.5} Retrievals from Multisource Data Fusion

Guannan Geng, Qingyang Xiao, Shigan Liu, Xiaodong Liu, Jing Cheng, Yixuan Zheng, Tao Xue, Dan Tong, Bo Zheng, Yiran Peng, Xiaomeng Huang, Kebin He, and Qiang Zhang*



Cite This: *Environ. Sci. Technol.* 2021, 55, 12106–12115



Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information

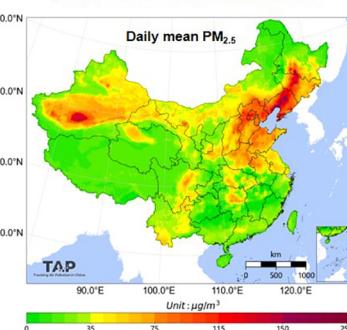
ABSTRACT: Air pollution has altered the Earth's radiation balance, disturbed the ecosystem, and increased human morbidity and mortality. Accordingly, a full-coverage high-resolution air pollutant data set with timely updates and historical long-term records is essential to support both research and environmental management. Here, for the first time, we develop a near real-time air pollutant database known as Tracking Air Pollution in China (TAP, <http://tapdata.org.cn/>) that combines information from multiple data sources, including ground observations, satellite aerosol optical depth (AOD), operational chemical transport model simulations, and other ancillary data such as meteorological fields, land use data, population, and elevation. Daily full-coverage PM_{2.5} data at a spatial resolution of 10 km is our first near real-time product. The TAP PM_{2.5} is estimated based on a two-stage machine learning model coupled with the synthetic minority oversampling technique and a tree-based gap-filling method. Our model has an averaged out-of-bag cross-validation R^2 of 0.83 for different years, which is comparable to those of other studies, but improves its performance at high pollution levels and fills the gaps in missing AOD on daily scale. The full coverage and near real-time updates of the daily PM_{2.5} data allow us to track the day-to-day variations in PM_{2.5} concentrations over China in a timely manner. The long-term records of PM_{2.5} data since 2000 will also support policy assessments and health impact studies. The TAP PM_{2.5} data are publicly available through our website for sharing with the research and policy communities.

KEYWORDS: PM_{2.5}, random forest, air pollution, exposure assessment, satellite remote sensing

Multisource data

- Ground observations
- Satellite AOD
- Operational WRF/CMAQ
- Meteorological fields
- Land use data
- Population
- Elevation

Near real-time retrievals



1. INTRODUCTION

With rapid urbanization and economic growth, anthropogenic emissions of reactive gases, aerosols, and aerosol precursors are being emitted into the atmosphere, and these substances substantially changed the atmospheric composition. Consequently, worsening air pollution has altered the Earth's radiation balance, distressed the ecosystem and increased the risks of human morbidity and mortality.^{1,2} In particular, one of the major air pollutants in China is fine particulate matter (PM_{2.5}), which could cause serious health problems^{3,4} and reduce visibility.⁵ Hence, understanding the spatial and temporal variations of ambient PM_{2.5} concentrations constitutes the basis for research studies associated with air pollution, climate change, and environmental health. It follows, then, that a complete-coverage high-resolution PM_{2.5} data set with timely updates and historical long-term records is essential to support both scientific research and environmental management. A complete-coverage data set would allow us to obtain a full spatial picture of PM_{2.5} pollution and identify heavily polluted areas that cannot be shown by the ground monitoring stations. Full-coverage data set can also help to

avoid exposure misclassification in epidemiological studies.^{6,7} Moreover, real-time or near real-time updates of PM_{2.5} data would help us track substantial changes in air pollution during haze events or special times such as the coronavirus pandemic. It could also be linked to real-time or near real-time acute effects of air pollution such as asthma flare-ups, hospital admissions and premature deaths, and alert the population for real-time public health prevention. In addition, a historical long-term data set benefiting from a consistent methodology could support clean air policy assessments and chronic health impact studies.^{8–11}

Several data sources could provide information about PM_{2.5} pollution. Among them, ground measurements are the most accurate way to obtain ambient PM_{2.5} concentrations.

Received: March 22, 2021

Revised: August 4, 2021

Accepted: August 4, 2021

Published: August 19, 2021



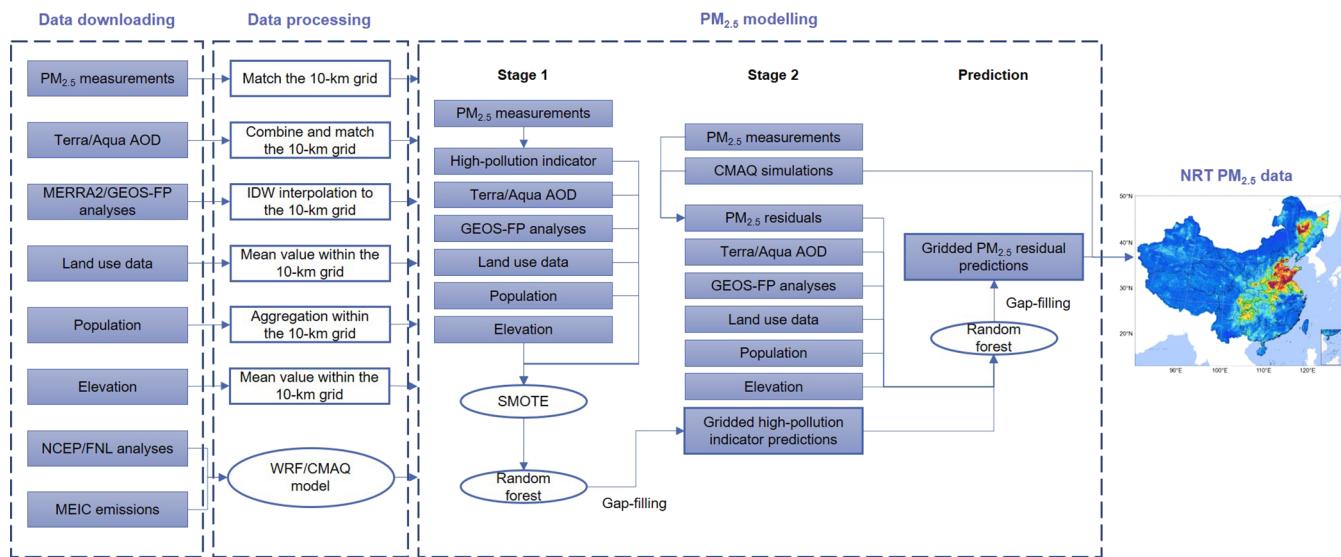


Figure 1. Operational process of the near real-time PM_{2.5} data generated from TAP.

Table 1. Summary of the Datasets Used in This Study from Multiple Sources

data category	data name	spatial resolution	temporal frequency	time coverage	data source
ground observations	PM _{2.5} measurements	point	hourly	2013 to date	http://www.cnemc.cn/
satellite AOD	MODIS Terra AOD	~10 km	daily	2000 to date	https://ladsweb.modaps.eosdis.nasa.gov/
	MODIS Aqua AOD	~10 km	daily	2000 to date	https://ladsweb.modaps.eosdis.nasa.gov/
operational WRF/ CMAQ	NCEP/FNL	1°	daily	2000 to date	https://rda.ucar.edu/datasets/ds083.2/
	NCEP/GFS	1°	daily	2000 to date	https://www.nco.ncep.noaa.gov/pmb/products/gfs/
	MEIC emissions	36 km	monthly	2000 to date	http://meicmodel.org/
	CMAQ simulations	36 km	daily	2000 to date	-
meteorological fields	MERRA-2	0.5°×0.625°	3-hly	2000 to date	https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
	GEOS-FP	0.5°×0.625°	6-hly	2011 to date	https://gmao.gsfc.nasa.gov/GMAO_products/NRT_products.php
land use data	FROM-GLC	30 m	yearly	2000–2018	http://data.ess.tsinghua.edu.cn/
population	GPW v4	1 km	yearly	2000 to date	https://beta.sedac.ciesin.columbia.edu/
	WorldPop	county	yearly	2000 to date	https://www.worldpop.org/
	China City Yearbooks	national	yearly	2000 to date	https://data.cnki.net/
elevation	GDEM	30 m	-	-	https://earthexplorer.usgs.gov/

However, due to the installation and maintenance costs of ground networks, monitoring stations are usually sparse and unevenly distributed, with most of the stations located in urban areas.^{11,12} Moreover, PM_{2.5} ground networks in China were established in 2013, so prior data are unavailable. As an alternative, chemical transport models (CTMs) could provide complete-coverage simulations of PM_{2.5} concentrations and could reproduce the spatial and temporal trends of PM_{2.5} concentrations when using reasonable emission inventories.^{10,13} However, biases still exist in the simulated absolute values of PM_{2.5} due to uncertainties in emission inventories¹⁴ and the lack of certain physical and chemical processes in the model.^{15–17} Satellite-retrieved aerosol optical depth (AOD) data, which can reflect the aerosol abundance in the atmosphere, have the advantage of long-term records and a high resolution. However, AOD data are missing during haze events, on cloudy days, and over bright surfaces such as desert and areas covered with snow, and the spatial distribution of missing data is sometimes nonrandom, which might cause biases in the long term average value.¹⁸ Accordingly, a data set that combines data from multiple sources is needed to take

advantage of all available information and to meet the requirements of such a data set, namely, a high accuracy, a full spatial coverage, a long temporal span, and real-time updates.

Previous studies have developed different methods to fuse two or more of the above data sets with other ancillary data in China to improve the estimation of PM_{2.5}.^{19–28} These methods include CTM-based algorithms,^{26,27} physical models,²¹ statistical models such as linear mixed-effects models and generalized additive models,^{19,22} and machine learning models such as random forest and extreme gradient boosting.^{20,23–25,28} As a result, numerous researchers have developed historical data sets in China, for example, 10 km PM_{2.5} data between 2000 and 2016 by Xue et al.²⁵ and 1 km PM_{2.5} data between 2000 and 2018 by Wei et al.²³ However, only some of these studies fill the gaps in AOD data using CTM simulations on a daily scale and achieve full-coverage daily PM_{2.5} concentrations.^{24,25,28} And previous studies usually underestimate PM_{2.5} concentrations on highly polluted days (e.g., PM_{2.5} > 150 µg/m³) due to the small sample size of high-pollution cases and highly nonlinear relationship between PM_{2.5} and

AOD.^{25,29} Furthermore, none of these works provide near real-time PM_{2.5} data publicly to support real-time public health prevention. Consequently, a near real-time data set with gap-filled daily PM_{2.5} estimates in China that can be shared with the research and policy communities remains lacking.

In this study, we develop the Tracking Air Pollution in China (TAP, <http://tapdata.org.cn/>) database based on an operational CTM, a two-stage machine learning model and a gap-filling method.³⁰ Our goal is to combine information from multiple data sources and provide near real-time (i.e., one-day delay) PM_{2.5} data on a daily scale with complete coverage at a spatial resolution of 10 km since 2000 to support related studies and environmental management. TAP is fully integrated on the cloud-computing platform, which allows users to conveniently access all customized data products online. Due to the downloading, processing and modeling procedures, the PM_{2.5} data of the previous day are available to the public at approximately 9:00 AM Beijing time.

2. DATA AND METHODS

Figure 1 shows the modeling framework of the TAP PM_{2.5} data set, including input data obtained from multiple data sources, processing of the input data, and the models developed to fuse the multisource data and generate near real-time PM_{2.5} retrievals.

2.1. Multisource Input Data. Table 1 summarizes all the input data used in this study, including ground observations, satellite AOD, operational Weather Research and Forecasting/Community Multiscale Air Quality Modeling System (WRF/CMAQ), and other ancillary data such as meteorological fields, land use data, total population, and elevation. The PM_{2.5} measurements updated every hour are collected from the national air quality monitoring network (<http://www.cnemc.cn/>) in China, which includes ~1600 stations over China. Continuous identical data over 3 h are excluded, and then the daily mean PM_{2.5} is calculated only if at least 12 hourly measurements are available. The PM_{2.5} measurements are matched to the 10 km grid cells they fall in.

Moderate Resolution Imaging Spectroradiometer (MODIS) Collection 6 level 2 aerosol products³¹ from both Aqua and Terra at a spatial resolution of 0.1° are downloaded from the National Aeronautics and Space Administration (NASA, <https://ladsweb.modaps.eosdis.nasa.gov/>) of the United States. We use AOD measurements retrieved by both the Dark Target (DT) algorithm³¹ and the Deep Blue (DB) algorithm³² from Terra and Aqua to improve the spatial coverage of AOD data. A daily linear regression is first fitted between the DT and DB AOD when both are available for Terra and Aqua separately and then used to fill the missing AOD when only the DT AOD or the DB AOD is valid. Then, the AOD averaged between the DT and DB AOD is calculated for Terra and Aqua separately. Similarly, a second linear regression is fitted between the Terra and Aqua AOD to fill the missing AOD when only one of them is available. The average Terra and Aqua AOD is used to represent the daily aerosol loading.³³

The WRF/CMAQ modeling system is included in our work to provide daily PM_{2.5} simulations. We employ the National Center for Environmental Prediction Final Analysis (NCEP-FNL, <https://rda.ucar.edu/datasets/ds083.2/>) and the Global Forecast System (NCEP-GFS, <https://www.nco.ncep.noaa.gov/pmb/products/gfs/>) to drive the WRF model. Anthropogenic emissions are taken from the Multiresolution Emission

Inventory in China (MEIC, <http://meicmodel.org/>),^{34,35} which is updated in a timely manner using a bottom-up approach based on near real-time activity indicators.³⁶ More details about the dynamic emissions can be found in Zheng et al.,³⁶ and a more in-depth description of the WRF/CMAQ model is provided in Section 2.2.1. Daily PM_{2.5} simulations from the WRF/CMAQ model are interpolated into the 10 km grid using the inverse distance weighting (IDW) method.

The meteorological analysis data are taken from the Modern-Era Retrospective Analysis for Research and Applications Version 2 (MERRA-2) data set at a resolution of 0.5° × 0.625°³⁷ and the Goddard Earth Observing System Forward Processing (GEOS-FP) data set at a resolution of 0.25° × 0.3125°.³⁸ We utilize the following parameters extracted from the analysis data: surface albedo, surface pressure, surface incoming shortwave flux, surface net downward shortwave flux, surface net downward longwave flux, total incoming shortwave flux, total net downward shortwave flux, total latent energy flux, cloud area fraction for low clouds, total cloud area fraction, total column ozone, total column odd oxygen, bias-corrected total precipitation, total precipitable ice water, total precipitable water vapor, total precipitable liquid water, evaporation from turbulence, 2 m specific humidity, 2 m dew point temperature, 2 m air temperature, 2 m eastward wind (U wind), 2 m northward wind (V wind), 10 m U wind, 10 m V wind, 10 m wind speed, U wind at 500 hPa, V wind at 500 hPa, U wind at 850 hPa, V wind at 850 hPa, planetary boundary layer height, and snowfall. Daily averaged meteorological analysis data are interpolated into the 10 km grid using the IDW approach.

We also download the urban and rural land cover classification data at a spatial resolution of 30 m from Gong et al.³⁹ (<http://data.ess.tsinghua.edu.cn/>) and elevation data at a spatial resolution of 30 m from the Global Digital Elevation Model (GDEM) version 2 (<https://earthexplorer.usgs.gov/>). Population data at a spatial resolution of 1 km are taken from the Gridded Population of the World (GPW) version 4 data set and are calibrated using the WorldPop data set at the county level and the total population reported in China City Yearbooks. The land cover data and elevation data are averaged within the 10-km grid while the population data are summed within the 10-km grid.

2.2. Algorithm Description. **2.2.1. Operational WRF/CMAQ Modeling System.** The TAP database includes an operational WRF/CMAQ modeling system to provide daily PM_{2.5} simulations as one of the input data sources, which could improve the accuracy of PM_{2.5} estimations and fill the gaps caused by missing AOD data. The WRF model version 3.9.1 and CMAQ model version 5.2 (<https://www.cmascenter.org/cmaq/>) are used in our work. The simulation domain covers all of China with a horizontal resolution of 36 km. The vertical resolution is designed as 46 sigma levels from the ground surface to 100 hPa for WRF but only 28 vertical layers in CMAQ after the processing of the Meteorology-Chemistry Interface Processor. For the WRF model, the NCEP-FNL and NCEP-GFS data are used to provide the initial and boundary conditions, whereas the NCEP-GFS sea surface temperature reanalysis data and NCEP Automated Data Processing global observational weather data are used for analysis, observation, and soil nudging. The parametrization scheme follows Cheng et al.⁴⁰ with the Kain-Fritsch cumulus physics scheme version 2⁴¹ modified to the Grell-Freitas ensemble scheme.⁴² For the CMAQ model, we use the CB05 gas-phase mechanism with

Table 2. Model Performance Compared with Other Studies Developing National PM_{2.5} Datasets in China

	gap-filled	spatial resolution	temporal resolution	CV type	CV R ²	RMSE ($\mu\text{g}/\text{m}^3$)
Ma et al. ²²	no	10 km	daily (2004–2013)	10-fold CV	0.79	27.4
Fang et al. ¹⁹	no	10 km	daily (2013–2014)	10-fold CV	0.80	22.8
He and Huang ⁵³	no	3 km	daily (2015)	10-fold CV	0.80	18.0
Xiao et al. ²⁴	yes	10 km	daily (2013–2017)	10-fold CV	0.79	21.0
Xue et al. ²⁵	yes	10 km	daily (2000–2016)	by-year CV	0.61	27.8
Liang et al. ²⁰	yes	1 km	monthly (2000–2016)	10-fold CV	0.93	6.2
Wei et al. ²³	no	1 km	daily (2013–2018) monthly (2000–2018)	10-fold CV	0.86–0.90	10.0–18.4
Huang et al. ²⁸	yes	1 km	daily (2013–2019)	10-fold CV by-year CV	0.87–0.88 0.62	11.9–21.9 27.7
TAP PM _{2.5}	yes	10 km	daily (2000–current)	out-of-bag CV (individual years) spatial CV (individual years) by-year CV (hindcast model)	0.80–0.88 0.69–0.83 0.58	13.9–22.1 14.6–26.4 27.5

the CMAQv5.1 update and sixth-generation CMAQ aerosol mechanism (AERO6).

The dynamically updated anthropogenic emissions for mainland China are taken from the MEIC.^{34–36} Emissions for other Asian countries and regions are obtained from the MIX inventory.⁴³ Biogenic emissions are calculated by the Model of Emissions of Gases and Aerosols from Nature (MEGAN) version 2.1.⁴⁴ Sea salt and dust aerosol emissions are calculated online by the CMAQ model.

The simulated PM_{2.5} concentrations from our WRF/CMAQ model have been fully evaluated against ground measurements in our previous studies.^{10,40,45} Accordingly, the model performance statistics can meet the recommended performance criteria, and the simulated results have been used for policy assessment and health impact studies in China.^{10,40,45}

2.2.2. Two-Stage Machine Learning Model. A two-stage machine learning model coupled with the synthetic minority oversampling technique (SMOTE) developed in our previous study⁴⁶ is used to generate the TAP PM_{2.5} data, as presented in Figure 1. In the first stage, we define a high-pollution indicator to improve the PM_{2.5} estimations on highly polluted days, when PM_{2.5} are usually underestimated in statistical and machine learning models.^{23,28} This high-pollution indicator is calculated based on PM_{2.5} observation data and describes whether the PM_{2.5} observations at each location exceed the monthly mean by two standard deviations. As high-pollution events cover only 3.9% of our training data set, which hinders the model's ability to characterize the associations between high-pollution events and other predictors, we adopt the SMOTE technique to resample our data set and obtain a balance between high-pollution and normal samples. The resampled data set is then used to train the first-stage random forest model with all the input data except for the CMAQ simulations, after which the predicted full-coverage high-pollution indicator is passed to the second-stage model as one of the input data. In the second stage, we use the residuals between the PM_{2.5} measurements and the CMAQ PM_{2.5} simulations as the dependent variable to train the second-stage random forest model. The predicted residuals combined with the CMAQ simulations represent the final PM_{2.5} estimations.

Compared with the models presented in previous studies, our model has two major advantages. In the first stage, the SMOTE algorithm balances the uneven proportion of high-pollution and normal data, which could improve the model performance at high PM_{2.5} levels. In the second stage, using the

residuals between simulated and measured PM_{2.5} enhances the variability of the dependent data, which could enhance the responses of predictors to PM_{2.5} variations, thus improving the prediction accuracy. We design a sensitivity test model (Sens) without the SMOTE technique and using PM_{2.5} measurements as the dependent variable to show our model improvements.

2.2.3. Gap-Filling Method. Our previous study³⁰ evaluated different gap-filling strategies and proposed a binary tree-based algorithm coupled with WRF/CMAQ simulations to fill the gaps in missing AOD. As the missingness of satellite AOD are primarily related to meteorological conditions (e.g., cloudy, rainy days) and PM_{2.5} pollution (e.g., highly polluted days), the tree-based algorithm could directly predict missing PM_{2.5} by mining the relationship between availability status of satellite data, PM_{2.5} concentrations and other Supporting Information.⁴⁷ This method is robust at characterizing the spatial patterns of PM_{2.5} without generating artificially oversmoothed PM_{2.5} spatial distributions and is efficient for use in a near real-time data product.³⁰ In each step of our two-stage model, a dichotomous predictor defined by whether the satellite AOD is available is constructed as the cut point of the first layer of the decision tree. This predictor serves to build the associations between satellite AOD availability, PM_{2.5} concentration, and other supportive information, such as WRF/CMAQ simulations and meteorological conditions, and helps to fill the gaps in the final PM_{2.5} estimations.

2.3. Operational Process of the TAP PM_{2.5} Data. Figure 1 shows the operational process for generating the near real-time PM_{2.5} product in TAP, which includes three steps: data downloading, data processing and PM_{2.5} modeling. Data from multiple sources (summarized in Table 1) are routinely downloaded to the cloud-computing platform every day once they are available. As these data are at different temporal and spatial resolutions, they are processed to match the 10 km grid defined in our work, as described in Section 2.1. Typically, the downloading and processing of the multisource input data are finished around 5:10–6:20 AM Beijing time.

Models are trained using input data from different time periods to develop PM_{2.5} data from 2000 to date. For years when ground PM_{2.5} measurements are available (i.e., 2013–2020), individual models are developed for these years using input data within each year. For the hindcast of PM_{2.5} prior to 2013 when ground measurements are absent, a model trained with data set between 2013–2019 is developed and validated to provide robust hindcasting power.

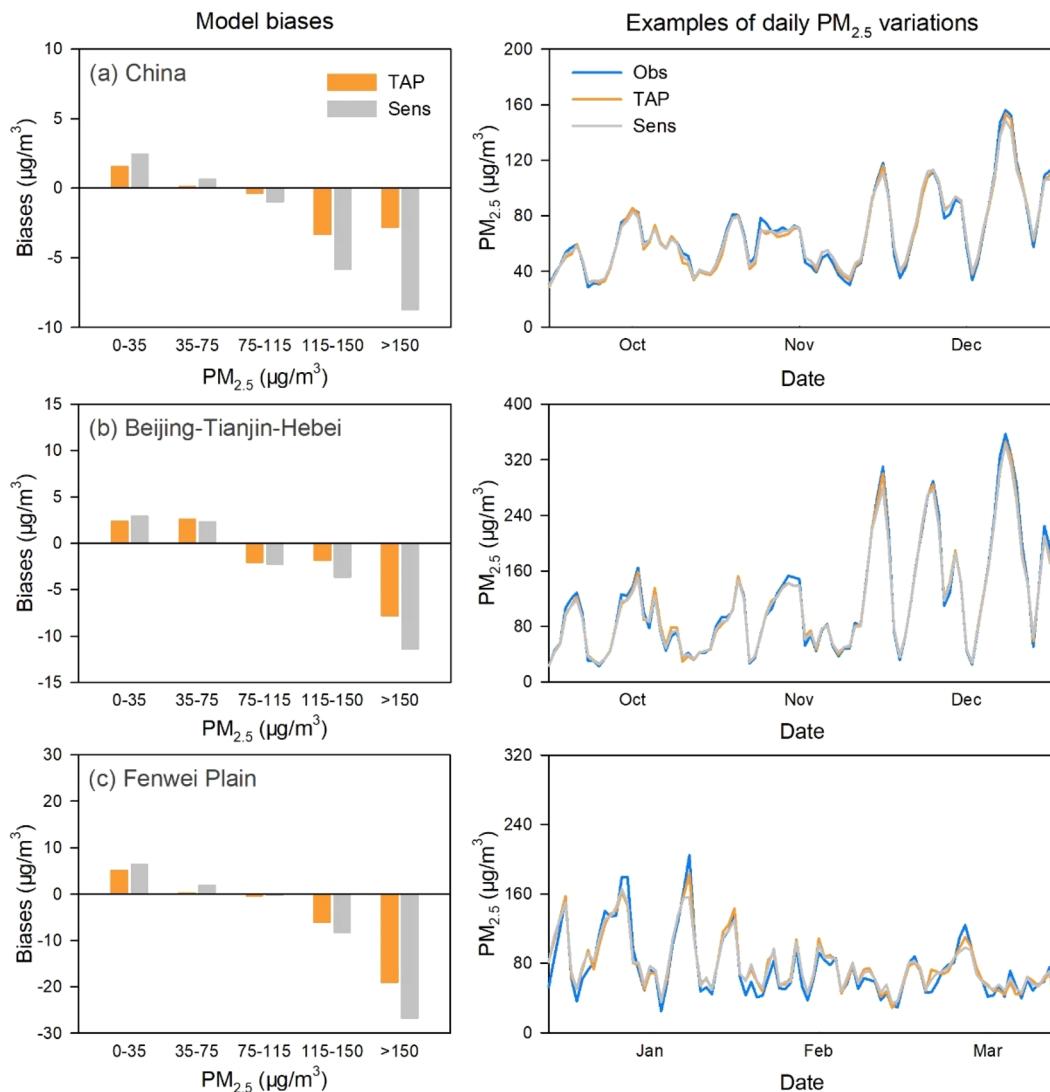


Figure 2. Comparison between the two-stage model in TAP and the sensitivity test model “Sens”. Left column: PM_{2.5} biases from TAP (orange) and Sens (gray) under different PM_{2.5} pollution levels in (a) China, (b) the Beijing-Tianjin-Hebei region and (c) the Fenwei Plain. Right column: examples of daily PM_{2.5} variations from ground observations (blue), TAP (orange), and Sens (gray).

For the near real-time product since Jan 2021, the training data set contains one-year data from the year 2020 and is updated every day on a rolling basis to include the most recent input data. The input data size (i.e., one year) is selected based on a series of sensitivity tests that use data of one, two, three and eight years to train the two-stage model (SI Table S1). When the size of the training data increases from one year to eight years, the time consumed to fit the model increases exponentially from hours to days (SI Table S1) while the model performance decreases slightly (SI Figure S1). Therefore, one-year data are selected in our near real-time model for both a reasonable fitting time and a good model performance. The two-stage random forest model is trained by the one-year rolling updated data set every day, and then near real-time PM_{2.5} data are generated and uploaded to our website. Usually, the daily PM_{2.5} prediction is generated no later than 8:25 AM Beijing time and uploaded to our website by 8:45 AM Beijing time.

3. EVALUATION OF MODEL PERFORMANCE

The performance of our two-stage model is evaluated through three cross-validation (CV) experiments: out-of-bag CV, spatial CV, and by-year CV. The out-of-bag CV is the most commonly used CV for the random forest models that compares the PM_{2.5} measurements with the predictions of out-of-bag samples. Spatial CV evaluates the model’s ability to make predictions at locations without monitors; all the monitoring stations are randomly divided into five subsets, and each time, the model is trained using data from four subsets and tested on the data from the remaining subset. Similarly, by-year CV evaluates the model’s hindcast prediction ability, which sequentially selects one year of data for testing and trains the model with the data from the remaining years.

Table 2 shows the CV results of our two-stage random forest models at the daily level, including the R^2 and root-mean-square error (RMSE) values between the CV estimates and the ground measurements. The PM_{2.5} predictions from the out-of-bag CV show good agreements compared against the ground observations, with R^2 of 0.80–0.88 and RMSE of 13.9–22.1 $\mu\text{g}/\text{m}^3$ for different years between 2013–2020, which indicates

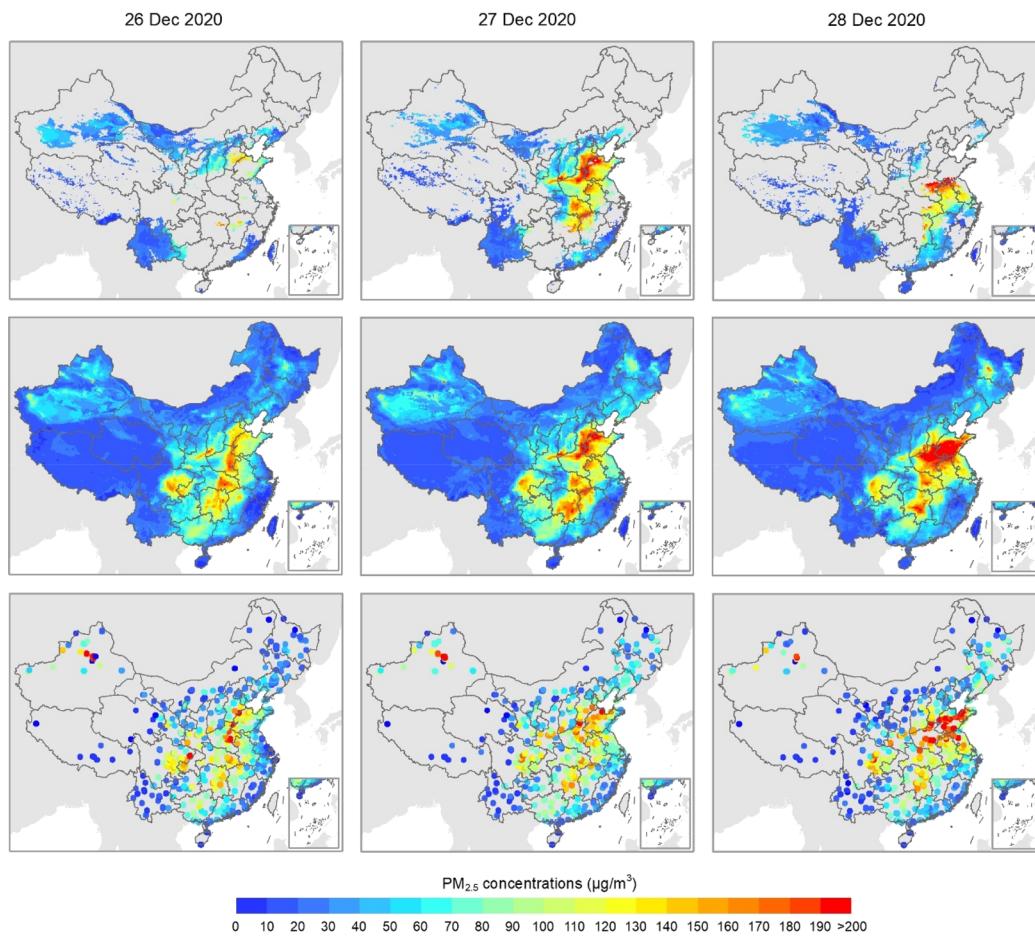


Figure 3. Daily full-coverage PM_{2.5} concentrations from TAP (middle row) compared with estimations without gap filling (top row) and the ground observations (bottom row). Data for 26–28 December 2020 are shown as examples.

that our two-stage model is quite robust. The spatial CV R^2 value decreases by 0.05–0.11 when compared with the out-of-bag CV, indicating that unobserved spatial trends contribute to the PM_{2.5} predictions. The model's hindcast performance further decreases in the by-year CV, with an R^2 of 0.58 and RMSE of 27.5 $\mu\text{g}/\text{m}^3$, reflecting a slight overfit in the hindcast of PM_{2.5} in years prior to 2013.

Our model's performance is comparable to that of models presented in other studies on the basis of the R^2 and RMSE values shown in Table 2. The statistical or machine learning models at the 10-km grid on a daily scale have 10-fold CV R^2 values ranging between 0.79 and 0.80 in China,^{19,22,24} which are similar to our out-of-bag CV results (i.e., 0.83 on average). Models with a 1 km grid have higher R^2 values,^{23,28} which might be partially explained by the correlations between PM_{2.5} and the 1 km AOD being higher than those between PM_{2.5} and the 10 km AOD, as well as the substantial increase in collocated AOD–PM_{2.5} pairs at a 1 km resolution than at a 10 km resolution for a larger sample size.⁴⁸

4. ILLUSTRATION OF TAP CAPABILITIES

4.1. Near Real-Time Updates. Our TAP PM_{2.5} product is the first near real-time PM_{2.5} database in China based on multisource data, including ground measurements, satellite AOD, high-resolution emission inventories (i.e., the MEIC inventory) and WRF/CMAQ simulations. Several factors support the timely update of PM_{2.5} data. First, the dynamic

updates of anthropogenic emissions in China by the MEIC and the high-performance computer at Tsinghua University facilitate the operational simulation of the WRF/CMAQ model, which is an important data source for PM_{2.5} estimations, as has been evaluated in previous studies.^{25,28} Second, we choose the tree-based algorithm to fill the gaps in PM_{2.5} concentrations, which is accurate and has reasonable speed. Other methods for filling in AOD gaps such as the multiple imputation method make use of more PM_{2.5} observations in the training data set;¹⁸ however, such a method has a much lower computation speed, and we found similar performances between these two gap-filling methods in our previous work.³⁰ Finally, the cloud-computing platform makes it possible to develop the model online and allows users to conveniently access all the data products. The daily data set of PM_{2.5} from TAP can be found through our website in near real time.

4.2. Improved Performance on Polluted Days. Our two-stage model coupled with the SMOTE technique improves the PM_{2.5} estimations on highly polluted days. Compared to the sensitivity test model (Sens) without SMOTE and using PM_{2.5} measurements directly as the dependent variable, the two-stage model has a similar R^2 but higher regression slope (0.97 vs 0.94) when evaluated against ground measurements. Figure 2 shows a detailed comparison between our two-stage model and the Sens model using year 2015 as an example. Usually, PM_{2.5} concentrations are underestimated over polluted days but a little overestimated

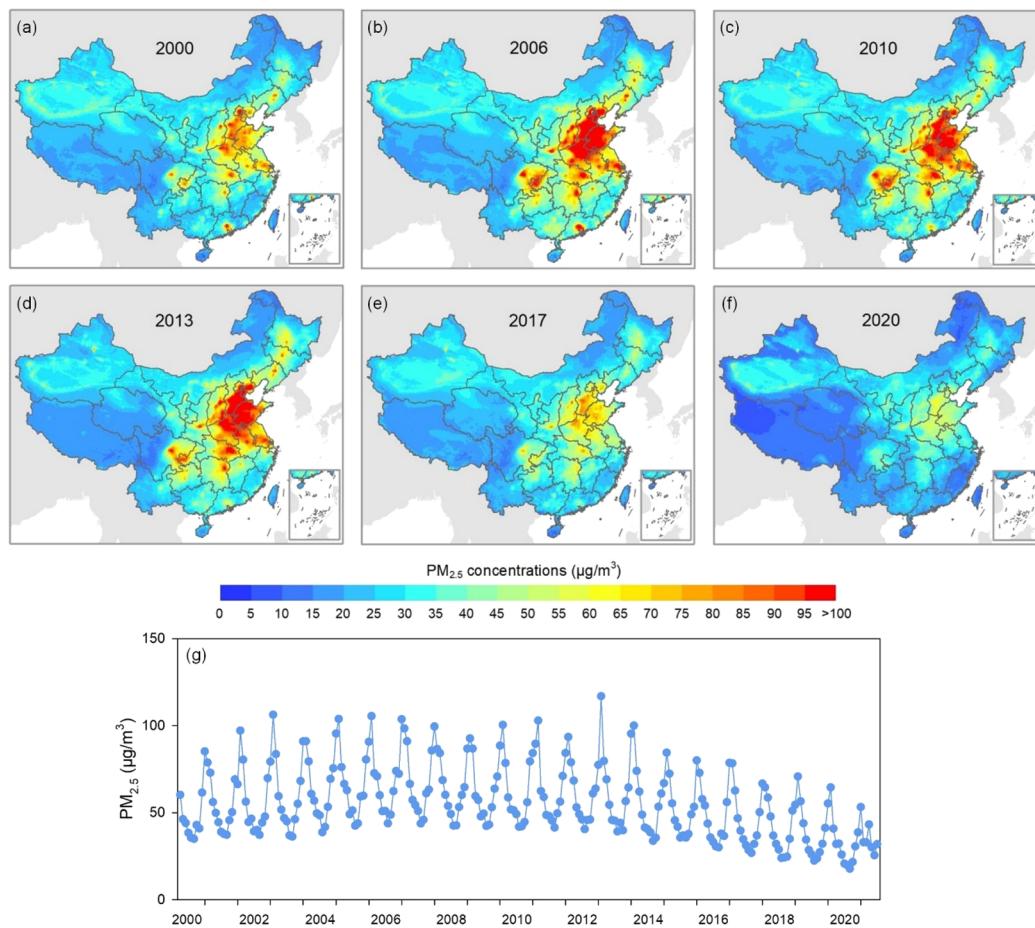


Figure 4. Historical data set of $\text{PM}_{2.5}$ from 2000 to the present. (a–f) Annual mean $\text{PM}_{2.5}$ concentrations for 2000, 2006, 2010, 2013, 2017, and 2020. (g) Population-weighted mean $\text{PM}_{2.5}$ in China from 2000 to the present on a monthly scale.

in clean days. After adopting our two-stage model, the mean biases over China at high $\text{PM}_{2.5}$ levels above $150 \mu\text{g}/\text{m}^3$ decrease by $5.9 \mu\text{g}/\text{m}^3$ (Figure 2a), and the number is even larger for the Fenwei Plain (i.e., $7.7 \mu\text{g}/\text{m}^3$). We also present examples of the estimated daily variations in $\text{PM}_{2.5}$ concentrations from TAP and Sens and find that TAP has better ability in capturing the concentrations peaks on polluted days.

4.3. Full-Coverage on Daily Scale. Figure 3 shows full-coverage daily maps of the TAP $\text{PM}_{2.5}$ as well as the estimations without gap-filling and ground observations during an example period, 26–28 December 2020. On these days, satellite AOD data only cover 16%–24% of the grid cells in China; thus, many $\text{PM}_{2.5}$ concentration hotspots are missing. Such missingness might cause biases in the averaged concentrations as the missing are sometimes nonrandom. In summer, AOD data are usually missing over southern China due to rain and clouds, while in winter, AOD data over northern China are usually missing due to snow cover and haze.⁴⁹ The nonrandom distribution of missing AOD causes negative biases in the average $\text{PM}_{2.5}$ in the north and positive biases in the average $\text{PM}_{2.5}$ in the south.^{18,26} After gap-filling with supportive information from the WRF/CMAQ simulations and meteorological data, the daily maps of $\text{PM}_{2.5}$ become complete and can more accurately capture the day-to-day variations in $\text{PM}_{2.5}$. For example, the TAP $\text{PM}_{2.5}$ maps successfully capture the $\text{PM}_{2.5}$ changes across the North China Plain during the haze event that occurred on 26–28 December

2020 (Figure 3). $\text{PM}_{2.5}$ pollution started to rise on 26 December 2020, and high $\text{PM}_{2.5}$ levels were found in southern Hebei and northern Shandong. Then, the pollution expanded, and on 28 December 2020, Shandong, Henan and northern Anhui were covered by haze exceeding $180 \mu\text{g}/\text{m}^3$. Such patterns could not be captured using the $\text{PM}_{2.5}$ data without gap filling.

4.4. Historical $\text{PM}_{2.5}$ Trends Since 2000. The TAP $\text{PM}_{2.5}$ database is also able to provide historical trends of $\text{PM}_{2.5}$ from 2000 to the present (Figure 4). Indeed, $\text{PM}_{2.5}$ estimates prior to 2013 have larger uncertainties, as there are no observation data to calibrate and evaluate our models. The by-year CV indicates that the model's hindcast ability has a smaller R^2 and larger RMSE than the out-of-bag CV. However, we use the year-by-year emission inventory from MEIC and the long-term CMAQ simulations as important input data to support the $\text{PM}_{2.5}$ estimates before 2013, thereby providing the best available knowledge of the spatial and temporal trends of $\text{PM}_{2.5}$ concentrations in history over China. Moreover, the long-term satellite AOD data set also provides valuable observational evidence of aerosol changes since 2000. We believe that the long-term trend of $\text{PM}_{2.5}$ constrained by these two data sets is reliable.

We also compare the $\text{PM}_{2.5}$ trends estimated by our data with the long-term $\text{PM}_{2.5}$ data set generated by Hammer et al.⁵⁰ (2000–2018) and the CHAP data developed by Wei et al.²³ (2000–2020) (SI Figure S2). Similar trends are found for the time when ground measurements are available, except for

CHAP in the pearl river delta. A possible reason is that CHAP is not gap-filled and the nonrandom missingness in AOD lead to biases in the annual mean PM_{2.5} value. For the time without available ground observations, the data by Hammer et al.⁵⁰ are always lower compared to TAP and CHAP. Emission estimates over China show that primary PM_{2.5} emissions and SO₂ emissions peaked around 2006,³⁴ which is more consistent with the trends estimated by TAP and CHAP.

Figure 4 shows the PM_{2.5} trends since 2000 in China. The TAP data capture the PM_{2.5} increase before 2006 (when there is no efficient emission control policy) and the sharp drop in PM_{2.5} concentrations after 2013 (when strict control measures were implemented). The peak of the national population-weighted mean PM_{2.5} concentrations occurred in 2006 (68.0 $\mu\text{g}/\text{m}^3$), the starting year of the 11th Five Year Plan (FYP, 2006–2010), when flue-gas desulfurization devices were installed in coal-fired power plants. After that, the increasing trend of PM_{2.5} concentrations was reversed. Since 2013, strict clean air policies have been implemented, that is, the Air Pollution Prevention and Control Action Plan (2013–2017) and the Blue Sky Protection Campaign (2018–2020). The Air Pollution Prevention and Control Action Plan reduced the annual population-weighted mean PM_{2.5} from 62.5 $\mu\text{g}/\text{m}^3$ in 2013 to 44.4 $\mu\text{g}/\text{m}^3$ in 2017, and the Blue Sky Protection Campaign further reduced the PM_{2.5} concentrations to 33.1 $\mu\text{g}/\text{m}^3$ in 2020.

5. DISCUSSION

In this study, we develop the TAP PM_{2.5} database that couples real-time ground observations, near real-time satellite data and meteorological reanalysis data, and operational simulations from the WRF/CMAQ modeling system to provide PM_{2.5} concentration data that are updated in a timely manner. Based on a two-stage machine learning model and gap-filling method, TAP provides daily full-coverage PM_{2.5} concentrations at a spatial resolution of 10 km in near real time. All the data are publicly available for sharing with the community.

Our work is subject to some limitations. First, our near real-time PM_{2.5} products rely on the near real-time updates of all the input data (except for the land use, population and elevation data, which have update frequencies of yearly or longer). Delays in any of these data sets would influence the updates of our PM_{2.5} data. Second, although we believe that the long-term spatial and temporal patterns of PM_{2.5} concentrations prior to 2013 are reliable due to the reasonableness of the input data, the uncertainties in PM_{2.5} on a daily scale are still larger than the daily PM_{2.5} estimates after 2013. Third, the TAP PM_{2.5} concentrations over the northwestern region dominated by the dust component tend to be underestimated, because MODIS AOD over bright surface (e.g., desert) are reported to have larger uncertainties compared to ground measurements^{51,52} and the CMAQ PM_{2.5} simulations underestimated dust concentrations severely.¹⁰ Finally, previous studies have shown that using 1 km AOD estimates from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm would improve the model performance, as a finer resolution would result in better correspondence between the AOD and PM_{2.5}.⁴⁸ However, building near real-time models at 1 km would cause exponential increases in the required computing resources and storage. Therefore, we choose the 10 km PM_{2.5} data as our first step for the TAP database.

In the future, we will continue to improve our methods and provide more air pollutant species and finer spatial resolution data. Accordingly, we will build the TAP database into a near real-time database of multiple air pollutants at different spatial and temporal resolutions based on multiple data sources.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.est.1c01863>.

Results of the sensitivity tests that aim at examining the impact of training data size on the model performance, comparisons of the estimated long-term PM_{2.5} trends during 2000–2020 between TAP and other data sets ([PDF](#))

AUTHOR INFORMATION

Corresponding Author

Qiang Zhang – Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China;
Email: qiangzhang@tsinghua.edu.cn

Authors

Guannan Geng – State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China; State Environmental Protection Key Laboratory of Sources and Control of Air Pollution Complex, Beijing 100084, China; [orcid.org/0000-0002-1605-8448](#)

Qingyang Xiao – State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China

Shigan Liu – Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

Xiaodong Liu – State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China

Jing Cheng – Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

Yixuan Zheng – Center of Air Quality Simulation and System Analysis, Chinese Academy of Environmental Planning, Beijing 100012, China; [orcid.org/0000-0002-3429-5754](#)

Tao Xue – Institute of Reproductive and Child Health, Ministry of Health Key Laboratory of Reproductive Health and Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; [orcid.org/0000-0002-7045-2307](#)

Dan Tong – Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China; [orcid.org/0000-0003-3787-0707](#)

Bo Zheng – Institute of Environment and Ecology, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

Yiran Peng – Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

Xiaomeng Huang — Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

Kebin He — State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China; State Environmental Protection Key Laboratory of Sources and Control of Air Pollution Complex, Beijing 100084, China

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.est.1c01863>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (42005135, 42007189, 41921005, and 41625020).

REFERENCES

- (1) Forouzanfar, M. H.; Afshin, A.; Alexander, L. T.; Anderson, H. R.; Bhutta, Z. A.; Biryukov, S.; Brauer, M.; Burnett, R.; Cercy, K.; Charlson, F. J. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **2016**, *388* (10053), 1659–1724.
- (2) Seinfeld, J. H.; Pandis, S. N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*; John Wiley & Sons, 2016.
- (3) Burnett, R.; Chen, H.; Szyszkowicz, M.; Fann, N.; Hubbell, B.; Pope, C. A.; Apte, J. S.; Brauer, M.; Cohen, A.; Weichenthal, S.; Coggins, J.; Di, Q.; Brunekreef, B.; Frostad, J.; Lim, S. S.; Kan, H.; Walker, K. D.; Thurston, G. D.; Hayes, R. B.; Lim, C. C.; Turner, M. C.; Jerrett, M.; Krewski, D.; Gapstur, S. M.; Diver, W. R.; Ostro, B.; Goldberg, D.; Crouse, D. L.; Martin, R. V.; Peters, P.; Pinault, L.; Tjeenkema, M.; van Donkelaar, A.; Villeneuve, P. J.; Miller, A. B.; Yin, P.; Zhou, M.; Wang, L.; Janssen, N. A. H.; Marra, M.; Atkinson, R. W.; Tsang, H.; Quoc Thach, T.; Cannon, J. B.; Allen, R. T.; Hart, J. E.; Laden, F.; Cesaroni, G.; Forastiere, F.; Weinmayr, G.; Jaensch, A.; Nagel, G.; Concin, H.; Spadaro, J. V. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (38), 9592–9597.
- (4) Cohen, A. J.; Brauer, M.; Burnett, R.; Anderson, H. R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; Feigin, V.; Freedman, G.; Hubbell, B.; Jobling, A.; Kan, H.; Knibbs, L.; Liu, Y.; Martin, R.; Morawska, L.; Pope, C. A.; Shin, H.; Straif, K.; Shaddick, G.; Thomas, M.; van Dingenen, R.; van Donkelaar, A.; Vos, T.; Murray, C. J. L.; Forouzanfar, M. H. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* **2017**, *389* (10082), 1907–1918.
- (5) Che, H.; Zhang, X.; Li, Y.; Zhou, Z.; Qu, J. J., Horizontal visibility trends in China 1981–2005. *Geophys. Res. Lett.* **2007**, *34*, (24). DOI: [10.1029/2007GL031450](https://doi.org/10.1029/2007GL031450)
- (6) Xiao, Q.; Chen, H.; Strickland, M. J.; Kan, H.; Chang, H. H.; Klein, M.; Yang, C.; Meng, X.; Liu, Y. Associations between birth outcomes and maternal PM_{2.5} exposure in Shanghai: a comparison of three exposure assessment approaches. *Environ. Int.* **2018**, *117*, 226–236.
- (7) Strickland, M. J.; Hao, H.; Hu, X.; Chang, H. H.; Darrow, L. A.; Liu, Y. Pediatric emergency visits and short-term changes in PM_{2.5} concentrations in the US State of Georgia. *Environ. Health Perspect.* **2016**, *124* (5), 690–696.
- (8) Geng, G.; Xiao, Q.; Zheng, Y.; Tong, D.; Zhang, Y.; Zhang, X.; Zhang, Q.; He, K.; Liu, Y. Impact of China's Air Pollution Prevention and Control Action Plan on PM_{2.5} chemical composition over eastern China. *Sci. China: Earth Sci.* **2019**, *62* (12), 1872–1884.
- (9) Geng, G.; Zhang, Q.; Tong, D.; Li, M.; Zheng, Y.; Wang, S.; He, K. Chemical composition of ambient PM_{2.5} over China and relationship to precursor emissions during 2005–2012. *Atmos. Chem. Phys.* **2017**, *17* (14), 9187.
- (10) Zhang, Q.; Zheng, Y.; Tong, D.; Shao, M.; Wang, S.; Zhang, Y.; Xu, X.; Wang, J.; He, H.; Liu, W.; Ding, Y.; Lei, Y.; Li, J.; Wang, Z.; Zhang, X.; Wang, Y.; Cheng, J.; Liu, Y.; Shi, Q.; Yan, L.; Geng, G.; Hong, C.; Li, M.; Liu, F.; Zheng, B.; Cao, J.; Ding, A.; Gao, J.; Fu, Q.; Huo, J.; Liu, B.; Liu, Z.; Yang, F.; He, K.; Hao, J. Drivers of improved PM_{2.5} air quality in China from 2013 to 2017. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 201907956.
- (11) Xiao, Q.; Geng, G.; Liang, F.; Wang, X.; Lv, Z.; Lei, Y.; Huang, X.; Zhang, Q.; Liu, Y.; He, K. Changes in spatial patterns of PM_{2.5} pollution in China 2000–2018: Impact of clean air policies. *Environ. Int.* **2020**, *141*, 105776.
- (12) Bai, K.; Li, K.; Wu, C.; Chang, N. B.; Guo, J. A homogenized daily in situ PM_{2.5} concentration dataset from the national air quality monitoring network in China. *Earth Syst. Sci. Data* **2020**, *12* (4), 3067–3080.
- (13) Xing, J.; Mathur, R.; Pleim, J.; Hogrefe, C.; Gan, C.-M.; Wong, D.-C.; Wei, C.; Gilliam, R.; Pouliot, G. Observations and modeling of air quality trends over 1990–2010 across the Northern Hemisphere: China, the United States and Europe. *Atmos. Chem. Phys.* **2015**, *15* (5), 2723–2747.
- (14) Zhang, Q.; Streets, D. G.; Carmichael, G. R.; He, K.; Huo, H.; Kannari, A.; Klimont, Z.; Park, I.; Reddy, S.; Fu, J. Asian emissions in 2006 for the NASA INTEX-B mission. *Atmos. Chem. Phys.* **2009**, *9* (14), 5131–5153.
- (15) Zheng, B.; Zhang, Q.; Zhang, Y.; He, K. B.; Wang, K.; Zheng, G. J.; Duan, F. K.; Ma, Y. L.; Kimoto, T. Heterogeneous chemistry: a mechanism missing in current models to explain secondary inorganic aerosol formation during the January 2013 haze episode in North China. *Atmos. Chem. Phys.* **2015**, *15* (4), 2031–2049.
- (16) Wang, Y.; Zhang, Q.; Jiang, J.; Zhou, W.; Wang, B.; He, K.; Duan, F.; Zhang, Q.; Philip, S.; Xie, Y. Enhanced sulfate formation during China's severe winter haze episode in January 2013 missing from current models. *Journal of Geophysical Research: Atmospheres* **2014**, *119* (17), 10,425–10,440.
- (17) Baek, J.; Hu, Y.; Odman, M. T.; Russell, A. G. Modeling secondary organic aerosol in CMAQ using multigenerational oxidation of semi-volatile organic compounds. *Journal of Geophysical Research: Atmospheres* **2011**, *116*, D22.
- (18) Xiao, Q.; Wang, Y.; Chang, H. H.; Meng, X.; Geng, G.; Lyapustin, A.; Liu, Y. Full-coverage high-resolution daily PM_{2.5} estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sensing of Environment* **2017**, *199* (Supplement C), 437–446.
- (19) Fang, X.; Zou, B.; Liu, X.; Sternberg, T.; Zhai, L. Satellite-based ground PM_{2.5} estimation using timely structure adaptive modeling. *Remote Sensing of Environment* **2016**, *186*, 152–163.
- (20) Liang, F.; Xiao, Q.; Huang, K.; Yang, X.; Liu, F.; Li, J.; Lu, X.; Liu, Y.; Gu, D. The 17-y spatiotemporal trend of PM_{2.5} and its mortality burden in China. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (41), 25601–25608.
- (21) Lin, C. Q.; Liu, G.; Lau, A. K. H.; Li, Y.; Li, C. C.; Fung, J. C. H.; Lao, X. Q. High-resolution satellite remote sensing of provincial PM_{2.5} trends in China from 2001 to 2015. *Atmos. Environ.* **2018**, *180*, 110–116.
- (22) Ma, Z.; Hu, X.; Sayer, A. M.; Levy, R.; Zhang, Q.; Xue, Y.; Tong, S.; Bi, J.; Huang, L.; Liu, Y. Satellite-Based Spatiotemporal Trends in PM_{2.5} Concentrations: China, 2004–2013. *Environ. Health Perspect.* **2016**, *124* (2), 184–192.
- (23) Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality PM_{2.5} data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sensing of Environment* **2021**, *252*, 112136.
- (24) Xiao, Q.; Chang, H. H.; Geng, G.; Liu, Y. An ensemble machine-learning model to predict historical PM_{2.5} concentrations in

- China from satellite data. *Environ. Sci. Technol.* **2018**, *52* (22), 13260–13269.
- (25) Xue, T.; Zheng, Y.; Tong, D.; Zheng, B.; Li, X.; Zhu, T.; Zhang, Q. Spatiotemporal continuous estimates of PM_{2.5} concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. *Environ. Int.* **2019**, *123*, 345–357.
- (26) Geng, G.; Zhang, Q.; Martin, R. V.; van Donkelaar, A.; Huo, H.; Che, H.; Lin, J.; He, K. Estimating long-term PM_{2.5} concentrations in China using satellite-based aerosol optical depth and a chemical transport model. *Remote Sensing of Environment* **2015**, *166*, 262–270.
- (27) Kong, L.; Tang, X.; Zhu, J.; Wang, Z.; Li, J.; Wu, H.; Wu, Q.; Chen, H.; Zhu, L.; Wang, W.; Liu, B.; Wang, Q.; Chen, D.; Pan, Y.; Song, T.; Li, F.; Zheng, H.; Jia, G.; Lu, M.; Wu, L.; Carmichael, G. R. A Six-year long (2013–2018) High-resolution Air Quality Reanalysis Dataset over China base on the assimilation of surface observations from CNEMC. *Earth Syst. Sci. Data Discuss.* **2020**, *2020*, 1–44.
- (28) Huang, C.; Hu, J.; Xue, T.; Xu, H.; Wang, M., High-Resolution Spatiotemporal Modeling for Ambient PM_{2.5} Exposure Assessment in China from 2013 to 2019. *Environ. Sci. Technol.* **2021**.552152
- (29) Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; Liu, L.; Wu, H.; Song, Y. Improved 1-km resolution PM_{2.5} estimates across China using enhanced space-time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20* (6), 3273–3289.
- (30) Xiao, Q.; Geng, G.; Cheng, J.; Liang, F.; Li, R.; Meng, X.; Xue, T.; Huang, X.; Kan, H.; Zhang, Q.; He, K. Evaluation of gap-filling approaches in satellite-based daily PM_{2.5} prediction models. *Atmos. Environ.* **2021**, *244*, 117921.
- (31) Levy, R.; Mattoe, S.; Munchak, L.; Remer, L.; Sayer, A.; Patadia, F.; Hsu, N. The Collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech.* **2013**, *6* (11), 2989.
- (32) Hsu, N.; Jeong, M. J.; Bettenhausen, C.; Sayer, A.; Hansell, R.; Seftor, C.; Huang, J.; Tsay, S. C. Enhanced Deep Blue aerosol retrieval algorithm: The second generation. *Journal of Geophysical Research: Atmospheres* **2013**, *118* (16), 9296–9315.
- (33) Jinnagara Puttaswamy, S.; Nguyen, H. M.; Braverman, A.; Hu, X.; Liu, Y. Statistical data fusion of multi-sensor AOD over the continental United States. *Geocarto International* **2014**, *29* (1), 48–64.
- (34) Li, M.; Liu, H.; Geng, G.; Hong, C.; Liu, F.; Song, Y.; Tong, D.; Zheng, B.; Cui, H.; Man, H. Anthropogenic emission inventories in China: a review. *National Science Review* **2017**, *4* (6), 834–866.
- (35) Zheng, B.; Tong, D.; Li, M.; Liu, F.; Hong, C.; Geng, G.; Li, H.; Li, X.; Peng, L.; Qi, J.; Yan, L.; Zhang, Y.; Zhao, H.; Zheng, Y.; He, K.; Zhang, Q. Trends in China's anthropogenic emissions since 2010 as the consequence of clean air actions. *Atmos. Chem. Phys.* **2018**, *18* (19), 14095–14111.
- (36) Zheng, B.; Zhang, Q.; Geng, G.; Shi, Q.; Lei, Y.; He, K. Changes in China's anthropogenic emissions during the COVID-19 pandemic. *Earth Syst. Sci. Data Discuss.* **2021**, *2020*, 1–20.
- (37) Gelaro, R.; McCarty, W.; Suárez, M. J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C. A.; Darmenov, A.; Bosilovich, M. G.; Reichle, R.; Wargan, K.; Coy, L.; Cullather, R.; Draper, C.; Akella, S.; Buchard, V.; Conaty, A.; da Silva, A. M.; Gu, W.; Kim, G.-K.; Koster, R.; Lucchesi, R.; Merkova, D.; Nielsen, J. E.; Partyka, G.; Pawson, S.; Putman, W.; Rienecker, M.; Schubert, S. D.; Sienkiewicz, M.; Zhao, B. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30* (14), 5419–5454.
- (38) Lucchesi, R. File Specification for GEOS-5 FP (*Forward Processing*). **2013**.
- (39) Gong, P.; Li, X.; Zhang, W. 40-Year (1978–2017) human settlement changes in China reflected by impervious surfaces from satellite remote sensing. *Science Bulletin* **2019**, *64* (11), 756–763.
- (40) Cheng, J.; Su, J.; Cui, T.; Li, X.; Dong, X.; Sun, F.; Yang, Y.; Tong, D.; Zheng, Y.; Li, Y.; Li, J.; Zhang, Q.; He, K. Dominant role of emission reduction in PM_{2.5} air quality improvement in Beijing during 2013–2017: a model-based decomposition analysis. *Atmos. Chem. Phys.* **2019**, *19* (9), 6125–6146.
- (41) Kain, J. S. The Kain?Fritsch Convective Parameterization: An Update. *Journal of Applied Meteorology* **2004**, *43* (1), 170–181.
- (42) Grell, G. A.; Freitas, S. R. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.* **2014**, *14* (10), 5233–5250.
- (43) Li, M.; Zhang, Q.; Kurokawa, J.-i.; Woo, J.-H.; He, K.; Lu, Z.; Ohara, T.; Song, Y.; Streets, D. G.; Carmichael, G. R., MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP. *Atmos. Chem. Phys.* **2017**, *17*, (2).935
- (44) Guenther, A. B.; Jiang, X.; Heald, C. L.; Sakulyanontvittaya, T.; Duhl, T.; Emmons, L. K.; Wang, X. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geosci. Model Dev.* **2012**, *5* (6), 1471–1492.
- (45) Zheng, Y.; Xue, T.; Zhang, Q.; Geng, G.; Tong, D.; Li, X.; He, K. Air quality improvements and health benefits from China's clean air action since 2013. *Environ. Res. Lett.* **2017**, *12* (11), 114020.
- (46) Xiao, Q.; Zheng, Y.; Geng, G.; Chen, C.; Huang, X.; Che, H.; Zhang, X.; He, K.; Zhang, Q. Separating emission and meteorological contribution to PM_{2.5} trends over East China during 2000–2018. *Atmos. Chem. Phys.* **2021**, *2021*, 1–32.
- (47) Brokamp, C.; Jandarov, R.; Hossain, M.; Ryan, P. Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model. *Environ. Sci. Technol.* **2018**, *52* (7), 4173–4179.
- (48) Chudnovsky, A.; Tang, C.; Lyapustin, A.; Wang, Y.; Schwartz, J.; Kourakis, P. A critical assessment of high-resolution aerosol optical depth retrievals for fine particulate matter predictions. *Atmos. Chem. Phys.* **2013**, *13* (21), 10907–10917.
- (49) Van Donkelaar, A.; Martin, R. V.; Levy, R. C.; da Silva, A. M.; Krzyzanowski, M.; Chubarova, N. E.; Semutnikova, E.; Cohen, A. J. Satellite-based estimates of ground-level fine particulate matter during extreme events: A case study of the Moscow fires in 2010. *Atmos. Environ.* **2011**, *45* (34), 6225–6232.
- (50) Hammer, M. S.; van Donkelaar, A.; Li, C.; Lyapustin, A.; Sayer, A. M.; Hsu, N. C.; Levy, R. C.; Garay, M. J.; Kalashnikova, O. V.; Kahn, R. A.; Brauer, M.; Apte, J. S.; Henze, D. K.; Zhang, L.; Zhang, Q.; Ford, B.; Pierce, J. R.; Martin, R. V. Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998–2018). *Environ. Sci. Technol.* **2020**, *54* (13), 7879–7890.
- (51) Tao, M.; Chen, L.; Wang, Z.; Wang, J.; Che, H.; Xu, X.; Wang, W.; Tao, J.; Zhu, H.; Hou, C. Evaluation of MODIS Deep Blue Aerosol Algorithm in Desert Region of East Asia: Ground Validation and Intercomparison. *Journal of Geophysical Research: Atmospheres* **2017**, *122* (19), 10,357–10,368.
- (52) Tao, M.; Chen, L.; Wang, Z.; Tao, J.; Che, H.; Wang, X.; Wang, Y. Comparison and evaluation of the MODIS Collection 6 aerosol data in China. *Journal of Geophysical Research: Atmospheres* **2015**, *120* (14), 6992–7005.
- (53) He, Q.; Huang, B. Satellite-based mapping of daily high-resolution ground PM_{2.5} in China via space-time regression modeling. *Remote Sensing of Environment* **2018**, *206*, 72–83.