Science of the Total Environment 827 (2022) 154363

Contents lists available at ScienceDirect

Science of the Total Environment

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

# MERRA-2 $PM_{2.5}$ mass concentration reconstruction in China mainland based on LightGBM machine learning

Jinghui Ma [a,b,c], Renhe Zhang [a,d,*], Jianming Xu [b,c], Zhongqi Yu [b,c]

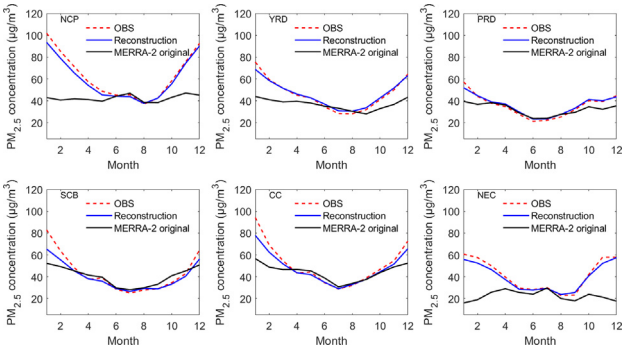
[a] Department of Atmospheric and Oceanic Sciences & Institute of Atmospheric Sciences, Fudan University, Shanghai 200438, China
[b] Shanghai Typhoon Institute, Shanghai Meteorological Service, Shanghai 200030, China
[c] Shanghai Key Laboratory of Meteorology and Health, Shanghai Meteorological Service, Shanghai 200030, China
[d] Big Data Institute for Carbon Emission and Environmental Pollution, Fudan University, Shanghai 200433, China

## HIGHLIGHTS

- A light gradient boosting machine (LGBM) model is developed to accurately estimate $PM_{2.5}$ concentrations in mainland China.
- A long-term $PM_{2.5}$ concentration across mainland China from 1980 to 2019 is reconstructed.
- The reconstructed daily, monthly, and annual $PM_{2.5}$ concentrations are highly accurate across mainland China.

## GRAPHICAL ABSTRACT

MERRA-2, reconstructed and observed monthly mean $PM_{2.5}$ concentrations in the six key regions.




## ARTICLE INFO

*Article history:*
Received 24 July 2021
Received in revised form 3 March 2022
Accepted 3 March 2022
Available online 6 March 2022

Editor: Pingqing Fu

*Keywords:*
MERRA-2
$PM_{2.5}$ reconstruction
Mainland China
Machine learning
Assessment

## ABSTRACT

MERRA-2 developed by the National Aeronautics and Space Administration (NASA) provides the long-term record of surface $PM_{2.5}$ mass concentration since 1980s, but needs great improvement over mainland China according to recent studies. In this study, a newly developed light gradient boosting machine (LGBM) model is introduced to correct the MERRA-2 $PM_{2.5}$ record over mainland China by incorporating the meteorological reanalysis and satellite AOD retrievals. A 40-year surface $PM_{2.5}$ record covering mainland China is reconstructed from 1980 to 2019, providing a new dataset for exploring the interactions between climate variability and air pollution. The new record exhibits not only much better magnitude but also more excellent variabilities of surface $PM_{2.5}$ loading compared to original MERRA-2 products. The correlation coefficient, the root-mean-square error and the mean error between the observed and reconstructed records are 0.8, less than 28.5 $\mu g \cdot m^{-3}$, and 0.33 $\mu g \cdot m^{-3}$, respectively, which are much better than those of 0.27, 45.8 $\mu g \cdot m^{-3}$, and 1.64 $\mu g \cdot m^{-3}$ between the observed and MERRA-2 $PM_{2.5}$ records. The $PM_{2.5}$ record with longer term and higher accuracy developed in this study provides a better base for the research on the climate change variability and air pollution in mainland China. However, limitations of the reconstructed record still exist, especially in the Tibetan Plateau and marine regions with very sparse surface measurements, which need further correction in the future studies.

## 1. Introduction

Fine particles related to economic development and climate change has become the major air pollutant in China in recent years (Bi et al., 2014; Mu and Zhang, 2014; Huang et al., 2015; Sun et al., 2016; Li et al., 2017a;

* Corresponding author at: Department of Atmospheric and Oceanic Sciences & Institute of Atmospheric Sciences, Fudan University, No. 2005 Songhu Road, Yangpu District, Shanghai 200438, China.
*E-mail address:* rhzhang@fudan.edu.cn (R. Zhang).

http://dx.doi.org/10.1016/j.scitotenv.2022.154363

Zhang, 2017). To meet the requirements of the air pollution monitoring, Chinese Atmosphere Watch Network (CAWNET) operated by Chinese Meteorological Administration (CMA) was initiated in 2006 to obtain the ambient aerosols on regional scale (Zhang et al., 2008; Guo et al., 2009; Wang et al., 2015). The national ambient air quality monitoring network, operated and maintained by the Ministry of Ecology and Environment of China, started in 2013 to monitor 6 types of major air pollutants. Such short time series of surface $PM_{2.5}$ measurements cannot meet the requirement of the research focusing on interactions of climate variability and air pollution.

The MERRA-2 products developed by the Global Modeling Assimilation Office (GMAO) (Randles et al., 2017) provide relative long-term surface $PM_{2.5}$ mass concentration since 1980s. However, Ma et al. (2020b) suggested that MERRA-2 underrates the $PM_{2.5}$ level across mainland China particularly in winter and autumn seasons. Buchard et al. (2016) pointed out that the bias in MERRA-2 mainly came from the uncertainties of the emission inventory and meteorological simulation in GOES-5 models. Meteorological condition plays an important role in the variability of air pollution in China (Zhang et al., 2014; Li et al., 2016; Wang and Zhang, 2020). Recent studies have shown the significant influence of annual meteorological variability on the $PM_{2.5}$ decrease by about 10–30% (Ding et al., 2019; Wang et al., 2019; Zhang et al., 2020). Wang et al. (2016) pointed out that the average concentration of air pollutants in winter of 2015 in China decreased by 20% compared with that in the same period of 2014, and about 40% of the decrease was possibly related to meteorological variations. The change of $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei (BTH) area in 2013–2018 was significantly controlled by meteorological conditions in both the interannual timescale (Wang and Zhang, 2020) and synoptic timescale (Wang et al., 2021). Even under the strict containment in anthropogenic activities during COVID-19 burst period in China, the meteorological factors could also lead to an increase in $PM_{2.5}$ around Beijing (Wang and Zhang, 2020b). It is also suggested that the meteorological factors can affect the interdecadal changes of $PM_{2.5}$ over China (Ma and Zhang, 2020). Therefore, it is expected that the meteorological factors should be carefully considered to improve the accuracy of both $PM_{2.5}$ levels and their variabilities by developing reconstructed datasets.

In the recent years, artificial intelligence, i.e., machine and deep learning, has been applied to construct datasets of near-surface $PM_{2.5}$ or $O_3$ concentrations, such as neural network (Di et al., 2017), random forest (RF) (Li et al., 2020; Wei et al., 2019; Wei et al., 2020; Wei et al., 2022), the fast space-time Light Gradient Boosting Machine (LGBM) (Wei et al., 2021a; Zhong et al., 2021), and Extreme Gradient Boosting (XGBoost) (Li et al., 2020; Chen et al., 2019). More and more estimated $PM_{2.5}$ concentrations were obtained (Chen et al., 2018; Li et al., 2017b; Wei et al., 2019; Wei et al., 2021b; Bai et al., 2021). These studies mostly provided gridded data of $PM_{2.5}$ concentrations with 1 km level resolution, allowing the efficient analysis on the $PM_{2.5}$ distribution and evolution in multi-scales, as well as the performance validation on the atmospheric chemistry models. However, those $PM_{2.5}$ datasets from above mentioned studies usually covered recent 5–20 years, which were insufficient for the research on the interactions of air pollution and climate variability. The dataset of surface $PM_{2.5}$ mass concentration with longer record and better accuracy needs to be developed.

The aim of this study is to develop the dataset of surface $PM_{2.5}$ mass concentration covering China mainland over 40 years based on the MERRA-2 products since 1980s. The LGBM model (Ke et al., 2017) is used to correct the MERRA-2 $PM_{2.5}$ bias in both spatial magnitude and temporal variabilities by integrating meteorological reanalysis and AOD products.

## 2. Data and methods

### 2.1. MERRA-2 products

The products of surface $PM_{2.5}$ mass concentration from MERRA-2 (Buchard et al., 2016; Buchard et al., 2017; Randles et al., 2017) were introduced in this study. The gridded MERRA-2 $PM_{2.5}$ records from 1980 to 2019 over China mainland were extracted day by day, as the original records for correction. We also consider the AOD of MERRA-2 as a factor in the LGBM model in this study.

### 2.2. Surface $PM_{2.5}$ measurements

The independent observations of $PM_{2.5}$ mass concentration were derived from the China ambient air quality monitoring network (https://air.cnemc.cn:18007/).There are 1534 ground stations in 2016, most of which distribute in the central and eastern China (Ma et al., 2020b). Daily averaged $PM_{2.5}$ observations from January 2014 to December 2019 are used for model training. The procedures from Song et al. (2018) are used for the quality control for the measurements. The surface measurements are interpolated to the grids with a resolution of $0.5^\circ \times 0.5^\circ$ by using the Kriging method (Beers and Kleijnen, 2003) to agree with the gridded MERRA-2 records. As an example, the spatial distributions of the $PM_{2.5}$ observational stations as well as the observed and MERRA-2 estimated annual mean $PM_{2.5}$ mass concentrations in 2014 and 2019 in mainland China are shown in Fig. S1 in Supplementary Materials.

### 2.3. Meteorological reanalysis data

Several previous studies have evaluated the robustness and accuracy of multiple reanalysis datasets of ERA5, MERRA-2, JRA55, and NCEP-2 in China (e.g., Guo et al., 2021; Bao and Zhang, 2019; Huang et al., 2021). In general, ERA5 has the best performance in China. The meteorological reanalysis data of ERA-5 is introduced as independent variables in this study for MERRA-2 $PM_{2.5}$ correction. Fifty-three meteorological factors, can be derived from ERA-5 reanalysis hourly data at the website https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form, including air temperature, relative humidity, total precipitation, U&V wind components etc. at surface and 42 vertical levels. Based on two wind components, the wind speed and wind direction are calculated by vector synthesis method.

### 2.4. Regions of interest (ROIs)

The key regions of top six urban clusters in China, namely, the North China Plain (NCP; 114°–120°E, 35°–41°N), the Yangtze River Delta (YRD; 118°–122°E, 29°–34°N), the Pearl River Delta (PRD; 112°–115°E, 22°–24°N), the Sichuan Basin (SCB; 103°–108°E, 28°–33°N), the central China (CC; 111°–116°E, 27°–32°N), and the Northeast China (NEC; 121°–127°E, 41°–48°N), are chosen in this study due to significant $PM_{2.5}$ issues exhibited in these regions in recent years. The locations of these six key regions are shown in Fig. S2 in Supplementary Materials. The number of observation stations and data efficiency in these regions is shown in Table 1.

### 2.5. Methods

The LGBM algorithm (Ke et al., 2017) is chosen for METTRA-2 $PM_{2.5}$ correction and reconstruction. LGBM is a decision tree machine learning algorithm. It contains two new techniques of exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS), which can deal with a great quantity of record occurrences and features. With the help of GOSS and EFB, LGBM can significantly outperform XGBoost (extreme gradient boosting algorithm) (Chen and Guestrin, 2016; Ma et al., 2020c; Zhong et al., 2021) in terms of computational speed and memory consumption by using a leaf-wise method with depth constraints to improve the

**Table 1**
Number of stations and efficiency of observations available in six key regions.

| Area | NCP | YRD | PRD | SCB | CC | NEC |
|---|---|---|---|---|---|---|
| Available station number | 196 | 556 | 180 | 98 | 130 | 114 |
| Effective rate of daily value | 95.1% | 96.2% | 95.3% | 91.3% | 92.7% | 93.7% |

accuracy of the model. This is more efficient than the level-wise method in XGBoost.

Three statistical metrics, including the correlation coefficient (R), the root-mean-square error (RMSE), and the mean error (ME), are applied to evaluate the LGBM performances (Zeng et al., 2020; Liu et al., 2019; Wang et al., 2017) by the 10-fold CV test for the surface $PM_{2.5}$ measurements. All the sampling data are randomly and evenly divided into 10 groups, nine groups for training and one for validation, which is repeated for 10 times. The results of 10-fold CV were averaged to get the final evaluate result. Fig. 1 shows the schematic process, which includes the following four steps:

**Step 1**. Data integration. The surface $PM_{2.5}$ measurements and ERA-5 reanalysis data are integrated to grids with 0.5° × 0.5° resolution, then both are normalized before being fed into the model.

**Step 2**. Parameters tuning. Scenario of 2016 was used for parameter tuning of the LGBM model to obtain optimal performance by testing different parameter combinations to obtain the best ones with the smallest residual error. The model performance with varying parameters is shown in Fig. S3 in the Supplementary Materials and the turned parameters for LGBM model are shown in Table 2. In order to guarantee the modeling accuracy, we also evaluated the tuned parameters by using the data in 2014 and 2019.

**Step 3**. Model training. The data series during 2014 and 2018 are selected for LGBM training to establish the equation for MERRA-2 $PM_{2.5}$ correction. The scenario of 2019 is used to evaluate the performance and effectiveness of the correcting method based on LGBM.

**Step 4**. Data Reconstruction. Finally, the daily MERRA-2 $PM_{2.5}$ records from 1980 to 2013 are corrected by the above mentioned LGBM method.

**Table 2**
Parameter tuning for LGBM model.

| Parameters[a] | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ |
|---|---|---|---|---|---|---|---|---|---|
| Values | 200 | 5 | 30 | 30 | 0.01 | 0.8 | 0.8 | 5 | default values |

[a] ① N_estimator, ② Max_depth, ③ Num_leaves, ④ Min_data_in_leaf, ⑤ Learning_rate, ⑥ Feature_fraction, ⑦ Bagging_fraction, ⑧ Bagging_freq, ⑨ other parameters.

To overcome the overfitting issue, the dataset enhancement (obtaining and using more data) and cross validation (10 CV to increases noise) are applied in this study. The selection procedure for the LGBM model used in this study is described in section 5 in Supplementary Materials.

## 3. Results and discussion

### 3.1. Model-fitting and validation

#### 3.1.1. Independent test of model performance

The performance of LGBM corrections is mainly evaluated in this section. For this reason, data of 2016 are chosen for model-fitting and validate the 2014 and 2019 predicted results based on the training model in the six key regions of mainland China. The density scatterplots of the fitted daily mean in 2016 for the six key regions are given in Fig. S4 in the Supplementary Materials. The LGBM model performs well, with R of 0.98–0.99 in the six key regions according to model-fitting daily mean results. The overall RMSE values are 1.7–5.4 $\mu g \cdot m^{-3}$ for the LGBM model in the six key regions on the basis of model-fitting daily mean results. This suggests that the LGBM model generates training approximations well.

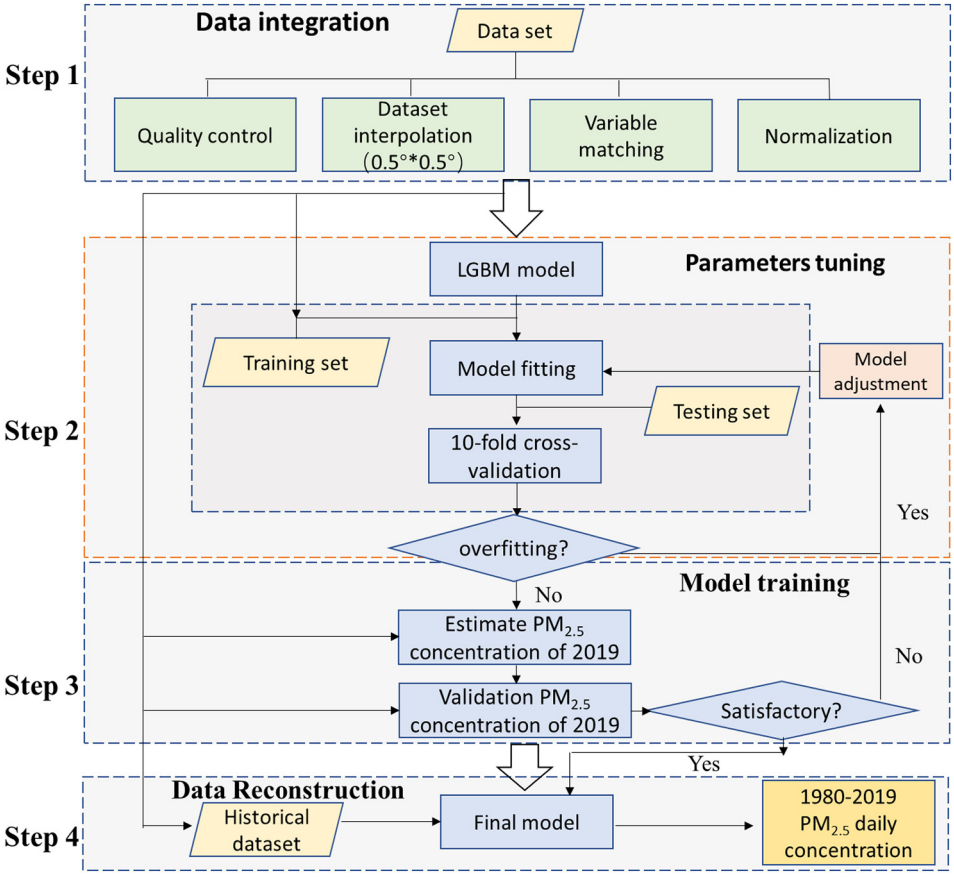



**Fig. 1.** Flowchart of MERRA-2 $PM_{2.5}$ estimation and reconstruction by LGBM model.

The daily-scale $PM_{2.5}$ estimates in 2014 produced by the 2016 data fitting LGBM model were close to observations (R = 0.5–0.79, and RMSE = 11.0–36.0 μg·m$^{-3}$ in the six key regions.) (Fig. 2a). Nevertheless, the LGBM model tended to underrate $PM_{2.5}$ concentrations in 2014 in the six key regions (slope = 0.3–0.64 and intercept is approximately 50.0 μg·m$^{-3}$), which may result in inferior estimates on severe polluted

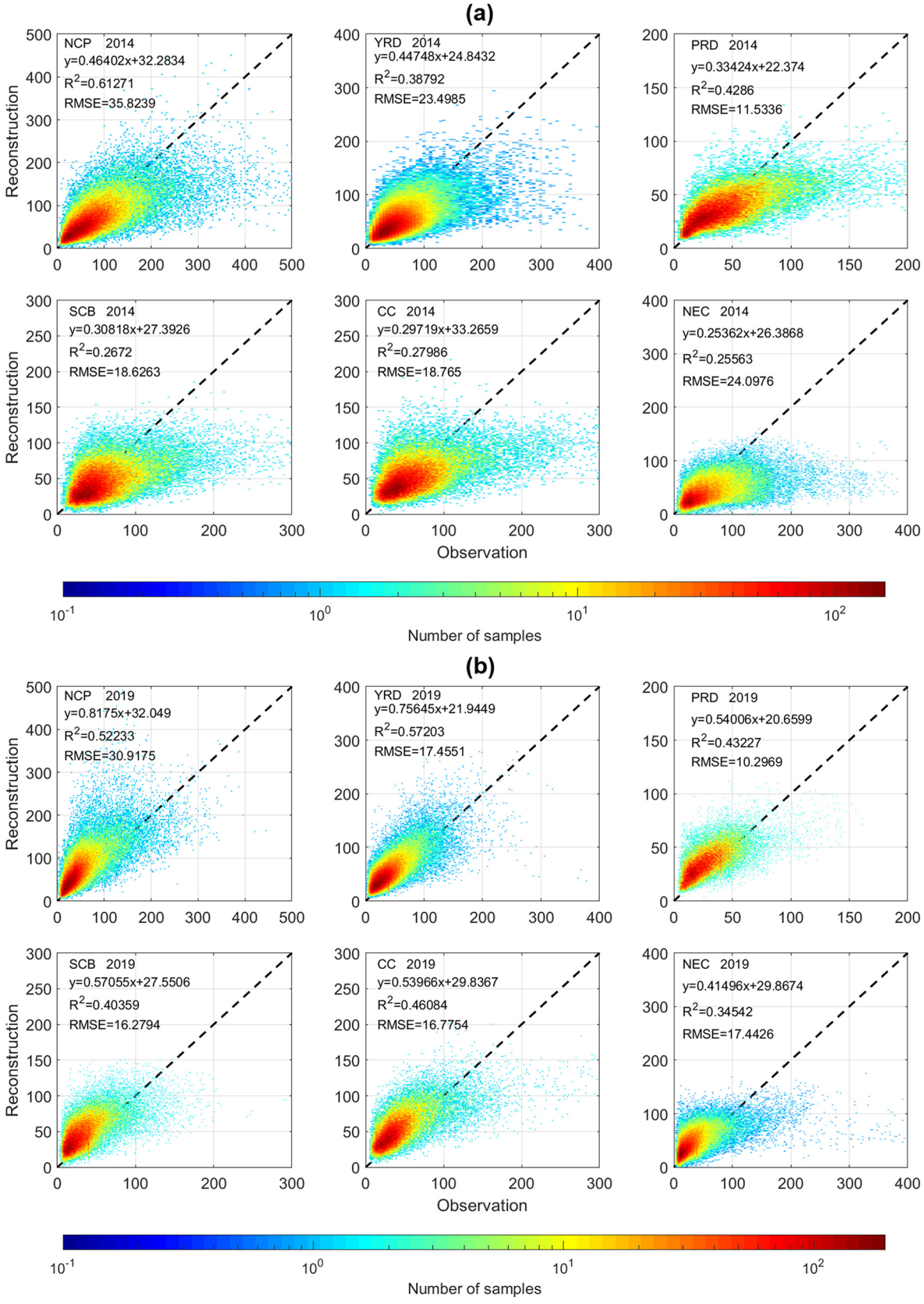



**Fig. 2.** Estimation and evaluation of predicted $PM_{2.5}$ daily mean concentration (unit: μg·m$^{-3}$) in 2014 and 2019 in the six key regions of China.

days. In 2019 the daily scale generated by the LGBM model was also close to measurements with R = 0.58–0.76 and RMSE = 10.0–31.0 μg·m$^{-3}$ in the six key regions (Fig. 2b). Compared with 2014, the fitting line of the reconstruction in 2019 was closer to the diagonal, indicating that the estimation ability of heavy pollution days is stronger. These results show the LGBM model can be applied well for the reconstruction of historical $PM_{2.5}$ concentration, which is highly significant in the study of long-term changes of $PM_{2.5}$ concentration and their causes in China.

### 3.1.2. 10-CV model performance

In the previous section, we conduct a systematic assessment of the predictive ability of LGBM model. We will use the 2014–2018 data to establish the LGBM model and 10-fold CV evaluation to determine the reliability of the LGBM model in this section.

The 10-fold CV of R, RMSE, and MB between the estimated and observed daily $PM_{2.5}$ in mainland China from 2014 to 2018 are shown in Fig. 3a–c. The R, RMSE, and MB were 0.6, 25.4 μg·m$^{-3}$ and 14.2 μg·m$^{-3}$, respectively. The spatial distributions of R, RMSE, and MB were inhomogeneous. Only a few areas in the northwest and northeast of mainland China had extremely high RMSE/MB values and low R values due to the relatively sparse $PM_{2.5}$ stations in these areas, which are difficult to provide enough training sets for model development. In central and eastern China, most of the R values exceeded 0.8 because of the dense $PM_{2.5}$ sites, providing enough samples to train and develop the model. In addition, the higher RMSE (>30 μg·m$^{-3}$) and MB (>25 μg·m$^{-3}$) values were mainly located in the NCP with higher industrial and anthropogenic emissions (Wei et al., 2019; Gui et al., 2019; Zhang et al., 2020), and northwestern China with higher natural dust emissions (An et al., 2018).

On a seasonal scale (Fig. 4), the LGBM model performed better in winter than the other three seasons in the three regions of the NCP, PRD, and NEC, with a steeper slope (0.54–0.77), increased R (0.81–0.93), and decreased RMSE (4.7–11.8 μg·m$^{-3}$). In the YRD, SCB, and CC regions, the model performed better in summer, with R being 0.89, 0.87, and 0.89, and RMSE 4.2, 3.1, and 2.9 μg·m$^{-3}$, respectively. However, the LGBM model tended to estimate $PM_{2.5}$ concentrations of daily-scale slopes between 0.54 and 0.78 for 10-fold CV in the six key regions, which led to poorer estimates on heavily polluted days. For the seasonal scale, underestimation also occurred, particularly in NEC, with the slopes of summer, autumn, winter, and spring were 0.42, 0.31, 0.45, and 0.36, respectively. The results of 10-fold CV confirm that the LGBM model performs well in estimating and predicting $PM_{2.5}$ concentrations in mainland China.

## 3.2. The model predictive ability

In the previous section, we evaluate the 10-fold CV results of the LGBM model and confirmed that the model can accurately estimate the $PM_{2.5}$ concentration in central and eastern China. In this section, we will use independent 2019 data to evaluate the model predictive ability trained by the data in 2014–2018.

Daily-scale density scatterplots of the 2019 forecast results by the LGBM model of China are shown in Fig. 5. The estimated results were close to the observations with R = 0.79–0.89 and RMSE = 7.5–16.5 μg·m$^{-3}$ in the six key regions. Compared to the other five regions, the error in the NEC was relatively large as a result of the large interpolation error due to sparse $PM_{2.5}$ stations, which lead to large errors of reconstructed $PM_{2.5}$ concentrations over there.

The MERRA-2, reconstructed and observed annual and seasonal $PM_{2.5}$ concentration distributions across China in 2019 are shown in Fig. 6. The annual and seasonal averages of $PM_{2.5}$ concentrations reconstructed by the LGBM model were very coincide with the observations in spatial distributions and values, particularly in winter. The annual averages of $PM_{2.5}$ concentrations of MERRA-2, observation and reconstruction in mainland China in 2019 were 30.6 ± 4.9 μg·m$^{-3}$, 38.6 ± 19.5 μg·m$^{-3}$, and 40.4 ± 18.0 μg·m$^{-3}$, respectively. The high $PM_{2.5}$ concentrations were mainly distributed in the NCP region and Xinjiang province. In the winter of 2019, the averaged $PM_{2.5}$ concentrations of MERRA-2, observation and reconstruction in mainland China were 51.3 ± 8.9 μg·m$^{-3}$, 68.8 ± 58.9 μg·m$^{-3}$ and 70.4 ± 47.8 μg·m$^{-3}$, respectively. High $PM_{2.5}$ concentrations were mainly distributed in the NCP, NEC, CC, YRD regions, Xinjiang and southern Shaanxi provinces. High $PM_{2.5}$ concentrations in these areas are mainly due to the rapid economic development and dense population, all of which contribute to the occurrence of high $PM_{2.5}$ concentrations. Notably, the high concentration of $PM_{2.5}$ in Xinjiang province is mainly attributed to the high anthropogenic aerosol emissions caused by heating in winter. The serious $PM_{2.5}$ pollution events in central and eastern China in winter are mainly caused by high concentrations of anthropogenic emissions from heating coal, and unfavorable diffusion conditions (Liao et al., 2018; Cai et al., 2017; Dang and Liao, 2019). Conversely, due to frequent precipitation, low anthropogenic emissions, favorable air pollution diffusion conditions, and other factors, the $PM_{2.5}$ concentration in summer usually presents a low value. It should be noted that there appeared unrealistic extreme high $PM_{2.5}$ over the Tibet Plateau, which may be caused by the interpolation deviation because of the sparse observations over there.

## 3.3. Model application: $PM_{2.5}$ seasonal hindcast and evaluation

Ma et al. (2020b) evaluated the MERRA-2 $PM_{2.5}$ concentration in mainland China and found that its seasonal variation in North China in the BHT is small. The largest bias was found in winter, reaching 50.4 μg·m$^{-3}$ and 40.3 μg·m$^{-3}$ in the NCP and NEC regions, respectively. This large bias is probably because of unresolved sources in the Goddard Chemistry, Aerosol, Radiation, and Transport model (GOCART). The $PM_{2.5}$ reconstruction concentration from 2014 to 2018 by the LGBM model is shown in Fig. 7. Compared with the MERRA-2 $PM_{2.5}$ concentration, the reconstruction had good agreement with observations and can describe the seasonal variation well. Comparing the reconstructions in the six key regions of the China mainland with MERRA-2, for example, the reconstruction values were closer to the observations, particularly the bias in winter reduced to

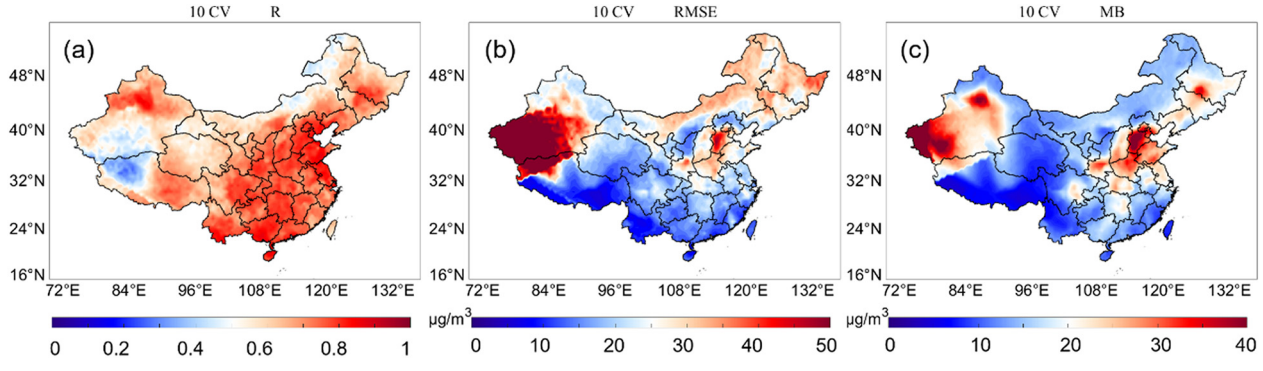



**Fig. 3.** Spatial distributions of the 10-fold cross validation (CV) of (a) R, (b) RMSE and (c) MB obtained from the reconstruction model from 2014 to 2018 over mainland China.

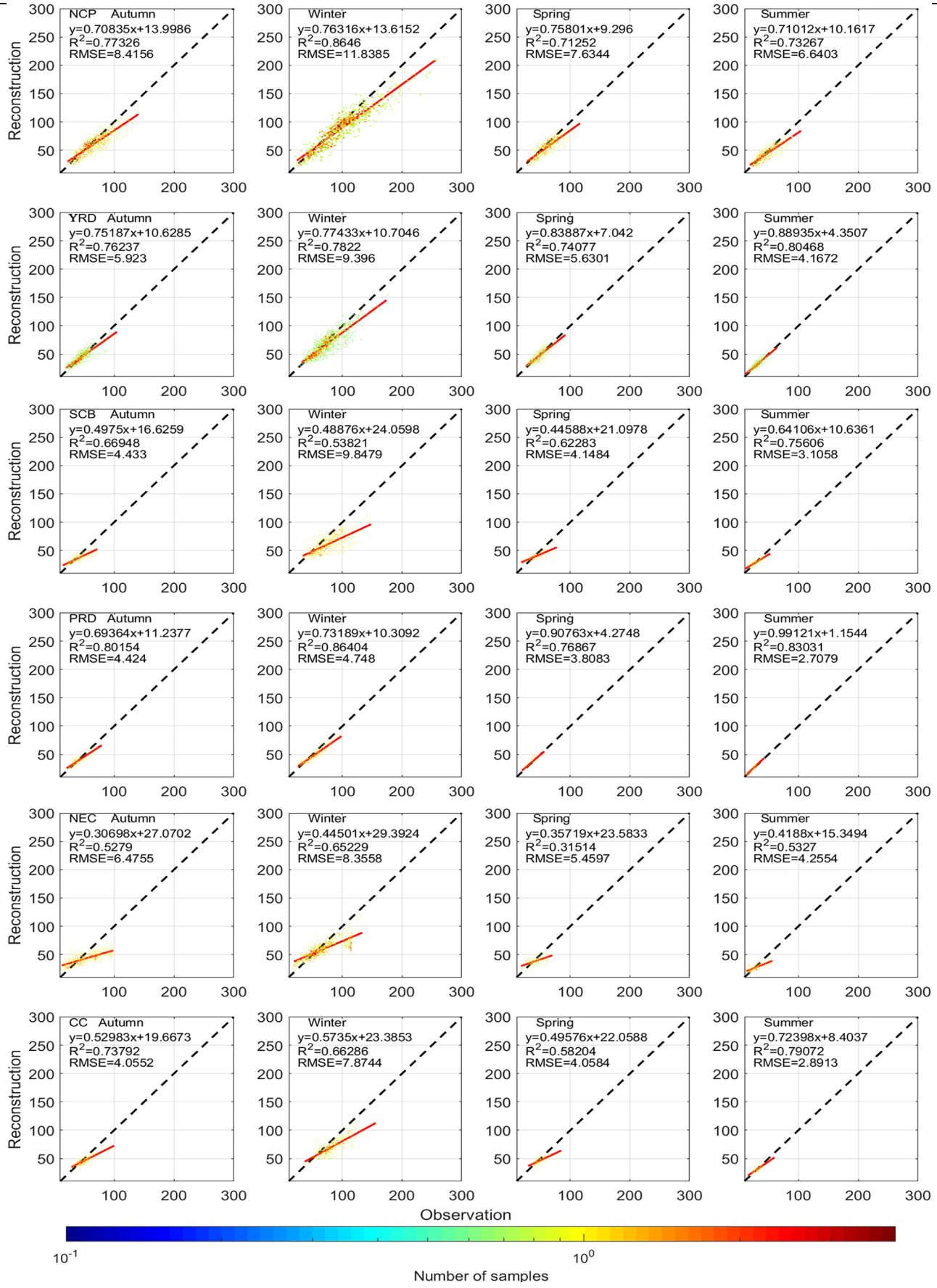



**Fig. 4.** Seasonal-scale density scatterplots of the 10-CV tests by the LGBM model from 2014 to 2018 in the six key regions of mainland China. The figures from top to bottom represent NCP, YRD, SCB, PRD, NEC and CC, respectively. The units of x-axis and y-axis are all $\mu g \cdot m^{-3}$.

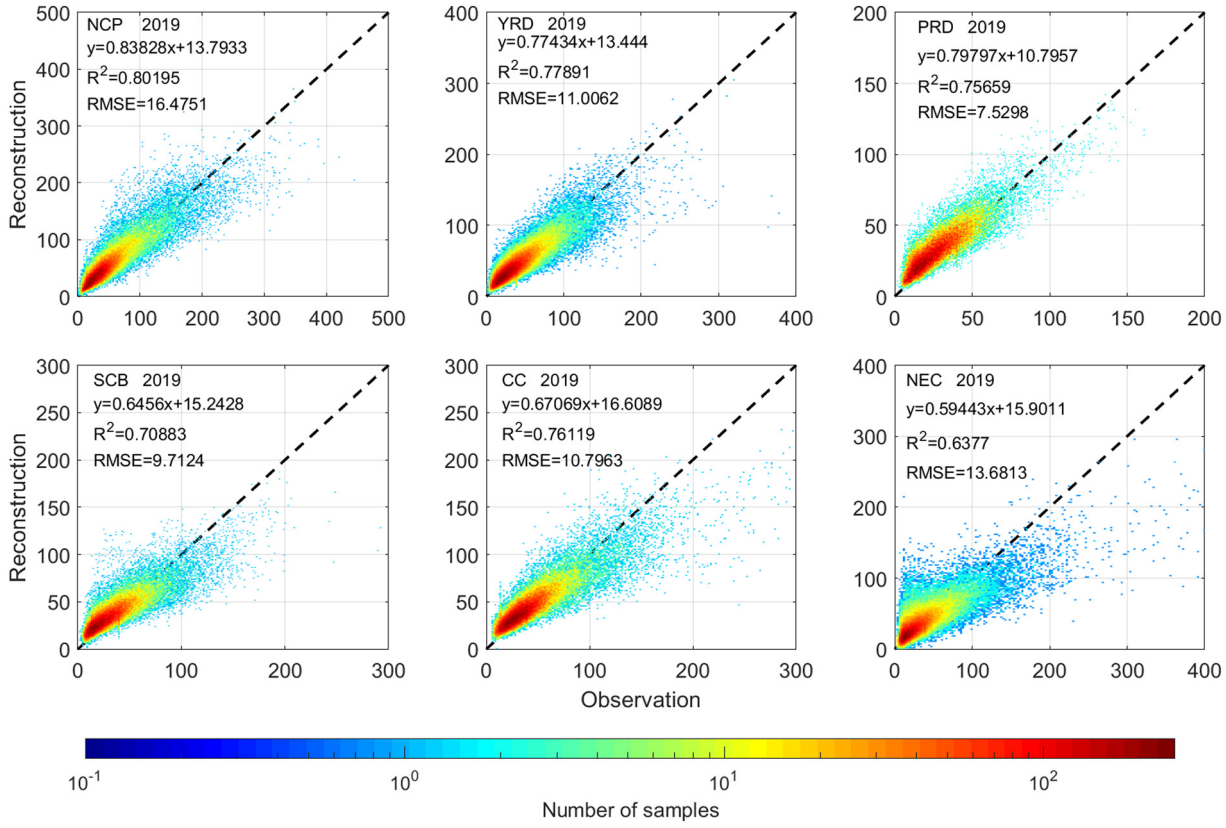



**Fig. 5.** Daily-scale density scatterplots of the 2019 prediction results by the reconstruction model in the six key regions of China. The units of x-axis and y-axis are all $\mu g \cdot m^{-3}$.

2.4 $\mu g \cdot m^{-3}$ and 0.4 $\mu g \cdot m^{-3}$ in the NCP and NEC regions, respectively. However, the correction of winter bias in southern China, such as in PRD and SCB regions, was small (approximately 5–10 $\mu g \cdot m^{-3}$), which is due to the difference between MERRA-2 and observations are noticeable smaller in these two regions than other regions.

Compared with the original value of MERRA-2, the reconstructed annual average of $PM_{2.5}$ concentrations by the LGBM model were closer to the observations from 2014 to 2019. The corrected annual averages of $PM_{2.5}$ concentrations in northern China (NCP and NEC) were 17.9 and 17.2 $\mu g \cdot m^{-3}$, respectively (Fig. 8). Furthermore, the revised values were 9.6 and 3.6 $\mu g \cdot m^{-3}$ in the YRD and CC, respectively, and the revised values were 4 and $-0.23$ $\mu g \cdot m^{-3}$ in PRD and SCB, respectively.

The LGBM model cannot effectively adjust the annual trend of MERRA-2. This is because in addition to meteorological factors, pollution source emissions also have a great impact on the $PM_{2.5}$ concentrations. The MERRA-2 emission source was set as the emissions of $SO_2$ derived from the Emissions Database for Global Atmospheric Research (EDGAR), version 4.2 database that covered the period 1970–2008 (European Commission, 2011). After this period, the assimilation system repeats the 2008 emissions. For primary sulfate ($SO_4$) aerosol emissions, the Aerosol Comparisons between Observations and Models (AeroCom) project covers the period 1979–2006 (http://aerocom.met.no/emissions.html). After this period, the assimilation system repeats the 2006 emissions. Therefore, during the 2014–2019 period, the anthropogenic emission source of MERRA-2 had almost no adjustment, and the LGBM model only corrects the part of the impact from meteorological factors on $PM_{2.5}$ concentration. Therefore, it cannot change the annual trend of the original $PM_{2.5}$ concentration.

### 3.4. Spatiotemporal variations of reconstructed data

Some studies explored the predictive power in estimating or reconstruction historical $PM_{2.5}$ concentrations by using different models, e.g., the Space-Time Extra-Trees (STET) model (Wei et al., 2020; Wei et al., 2021a), the space-time random forest (STRF) model (Wei et al., 2019; Hu et al., 2017), the two-stage model (Ma et al., 2016; Xue et al., 2019). Table 3 shows that the predicted historical $PM_{2.5}$ concentrations by these models were relatively poor at daily and monthly scales. Our LGBM model shows a relatively stronger predictive ability compared to those models developed in previous studies. These results suggest that our LGBM model can be more powerful and accurate in reconstructing a long-term historical $PM_{2.5}$ dataset.

Fig. 9 presents the temporal distributions of monthly mean $PM_{2.5}$ across the six key regions of China from 1980 to 2019 obtained from the MERRA-2 and the reconstruction. Compared with the MERRA-2, the predicted monthly average $PM_{2.5}$ is in good accord with the observed seasonal variation, being high in winter and low in summer. In general, the correction ranges of the reconstructed data to the seasonal variation in northern China, for example, NCP and NEC, were relatively large, and the average correction values were 40–45 $\mu g \cdot m^{-3}$ in winter (NDJ) and 1–10 $\mu g \cdot m^{-3}$ in summer (JJA), respectively. Furthermore, for the PRD and SCB, the average correction values were 5–10 $\mu g \cdot m^{-3}$ in winter and less than 5 $\mu g \cdot m^{-3}$ in summer. For the YRD and CC, the average corrections were 10–20 $\mu g \cdot m^{-3}$ in winter and 8–12 $\mu g \cdot m^{-3}$ in summer. When comparing the reconstructed data in the six key regions with the original data of MERRA-2, it is found that the variance of the reconstructed $PM_{2.5}$ concentration is smaller than that of the original MERRA-2 data. This indicates that the monthly mean reconstructions are more convergent to the mean value and the stability of the data is improved.

The reconstructed spatial patterns of annual and seasonal $PM_{2.5}$ mass concentrations are unevenly in China (Fig. 10), where the highest $PM_{2.5}$ values are found over NCP, CC, YRD areas and Xinjiang province, mainly due to economic development and urbanization, unique topographic conditions, and frequent dust events. Moreover, a strong north-to-south decreasing gradient is found, which agrees with the findings of previous studies (Lin et al., 2015; Wei et al., 2021b). The spatial patterns of $PM_{2.5}$

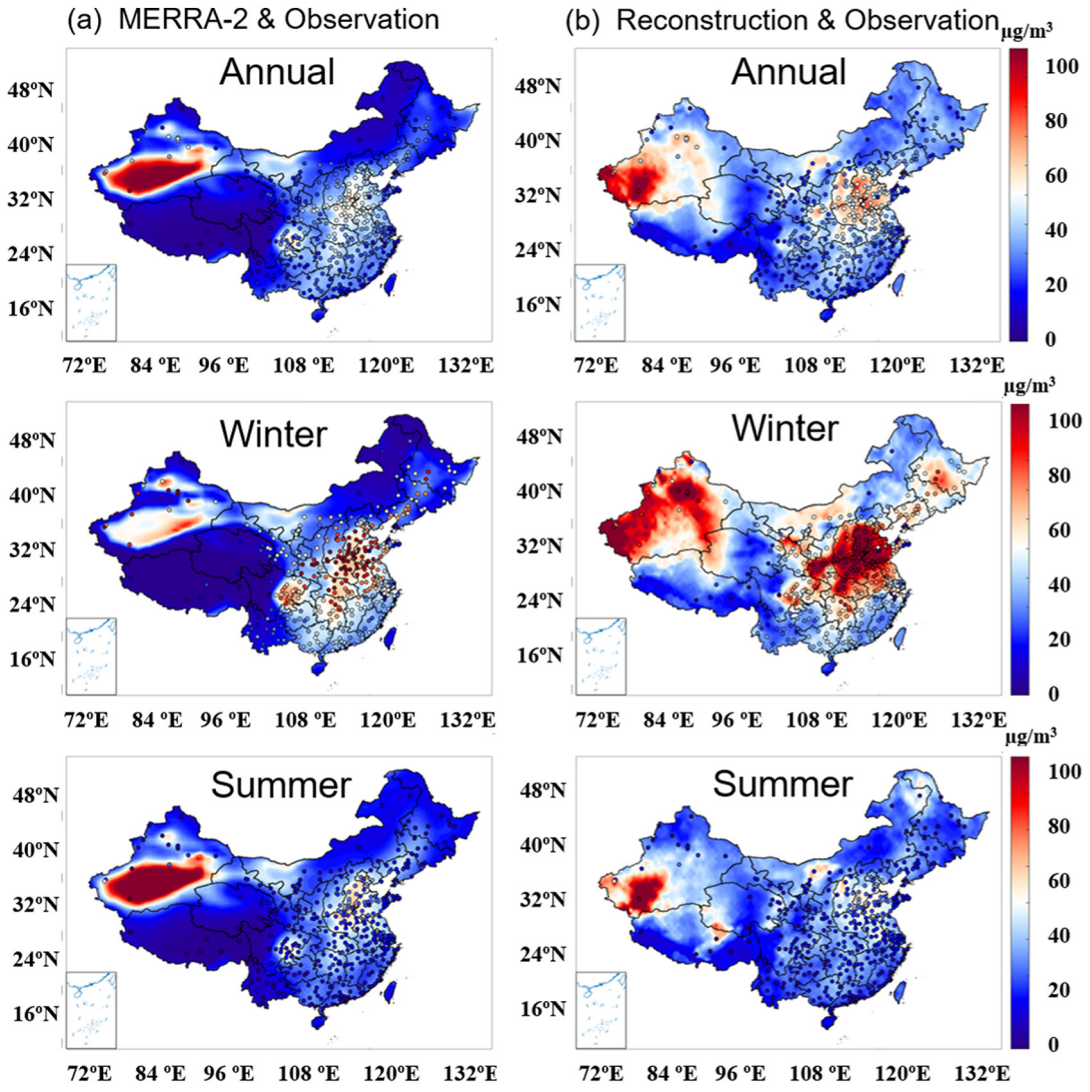



**Fig. 6.** Spatial distributions of annual and seasonal mean $PM_{2.5}$ from (a) MERRA-2, (b) the reconstruction model (shadings), and the CNEMC network (colored dots) in 2019. The figures from top to bottom represent the annual, winter (NDJ, November–December–January), and summer (JJA, June–July–August), respectively.

concentration greatly differ at the seasonal level. In summer, $PM_{2.5}$ pollution is the lightest, with most areas in China having $PM_{2.5}$ values <40 $\mu g \cdot m^{-3}$. By contrast, $PM_{2.5}$ pollution is the most severe in winter, with much of China having $PM_{2.5}$ values >70 $\mu g \cdot m^{-3}$. Except for Xinjiang province, where dust events frequently occur, spring and autumn have similar spatial patterns and pollution levels from regional to national scales. The seasonal distribution characteristics of $PM_{2.5}$ are also consistent with the previous studies (Lin et al., 2015; Wei et al., 2021b).

Fig. 11 shows the temporal distributions of reconstructed annual averages of $PM_{2.5}$ in the six key regions of China from 1980 to 2019. The reconstructed $PM_{2.5}$ concentrations in the six key regions of mainland China have obvious regional differences. In general, the annual concentration of $PM_{2.5}$ was the highest in NCP and the lowest in PRD. According to the year-to-year changes of the reconstructed $PM_{2.5}$, the trends in the six key regions can be divided into three stages. The 1980–2000 period was a slow rising stage, generally maintaining a low value, the 2000–2013 period a stage of rapid rise, and 2014–2019 the stage of overall decline. It is well known that pollution emission is an important factor affecting the trend of pollution concentration. Our LGBM model only considers the influence of meteorological factors and does not consider the change in pollution emission sources. Therefore, the LGBM model can correct the $PM_{2.5}$ concentration of MERRA-2 but cannot change the trend.

The increasing trend of $PM_{2.5}$ concentration around 2000 in China was also appeared in the previous studies (e.g., Wei et al., 2021a; Xue et al., 2019), in which the $PM_{2.5}$ was estimated by the satellite AOD data. However, the increasing was not so sharp as appeared in our study. China's economy has developed rapidly since 1980, and the interannual variation of $PM_{2.5}$ is greatly affected by anthropogenic emission sources. MERRA-2 used the AeroCom Phase II dataset (HCA0 v1) described in Diehl et al. (2012) (http://aerocom.met.no/emissions.html). Anthropogenic emissions of $SO_2$ derive from the EDGAR, version 4.2 database that covers the period 1970–2008 and after this period, it repeats the 2008 emissions. Therefore, the trend of $PM_{2.5}$ concentration in MERRA-2 after 2008 only depends on the influence of meteorological factors. This indicates that the influence of meteorological factors on $PM_{2.5}$ concentration is highly significant, and is agreement with the results of previous research (Zhang et al., 2014; Zhang, 2017; Zhang et al., 2019).

## 4. Conclusions and discussions

In this study, we utilize LGBM model reconstruct a $PM_{2.5}$ dataset covering mainland China by using MERRA-2 reanalysis $PM_{2.5}$ concentration data, AOD reanalysis data, and ERA-5 meteorological data. This study aims at solving the problems that the observed $PM_{2.5}$ mass concentration

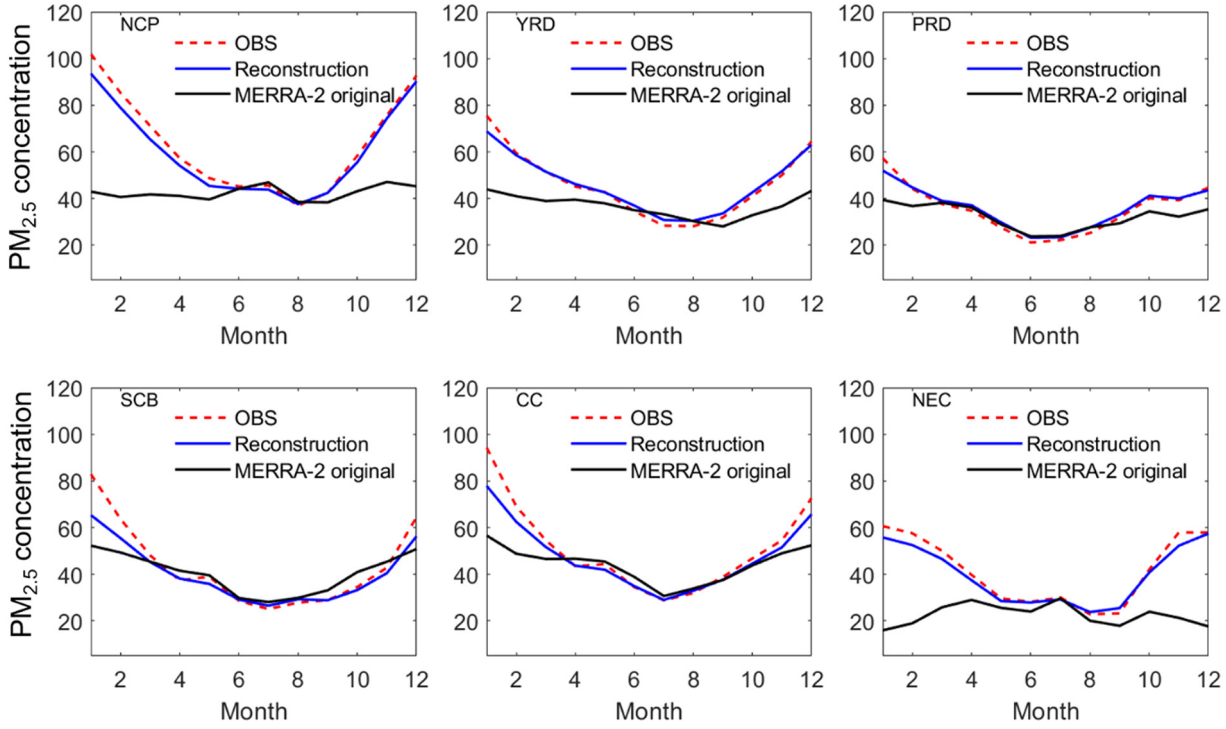



**Fig. 7.** MERRA-2, reconstructed and observed monthly mean $PM_{2.5}$ concentrations (unit: $\mu g \cdot m^{-3}$) in the six key regions from 2014 to 2019.

time series in China is relatively short (from 2013 to present), and the MERRA-2 $PM_{2.5}$ mass concentration in mainland China is significantly low (Ma et al., 2020b). There is an urgent need to reconstruct a set of long period and accurate $PM_{2.5}$ concentration data for pollution climate research and prediction.

In the present study, we first interpolate the $PM_{2.5}$ concentration data at 1460 observation sites across China into 0.5° × 0.5° grid points by using Kriging and nearest neighbor methods. The data of the MERRA-2 $PM_{2.5}$ concentration reanalysis are then interpolated from a 0.5° × 0.625° grid points to 0.5° × 0.5°grid points by the same method to keep consistency with the grid resolution of ERA-5 meteorological reanalysis data. Based on the LGBM model, a long-term (1980–2019) $PM_{2.5}$ concentration dataset with 0.5° × 0.5° resolution in mainland China is reconstructed. Taking the NCP region as an example, the R and RMSE values of estimated daily (monthly) $PM_{2.5}$ concentrations in 2014 were 0.78 (0.79) and 35.8 $\mu g \cdot m^{-3}$ (16.3 $\mu g \cdot m^{-3}$), respectively, which are highly consistent

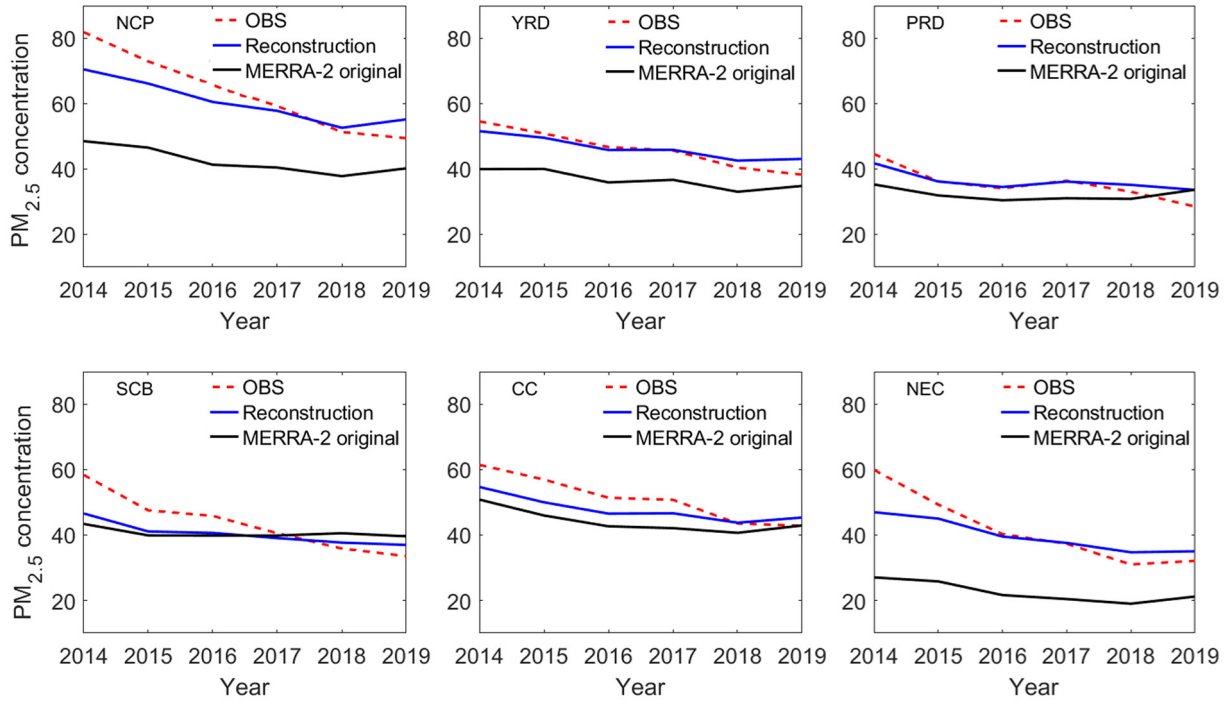



**Fig. 8.** Annual mean $PM_{2.5}$ concentrations (unit: $\mu g \cdot m^{-3}$) in the six key regions from 2014 to 2019.

**Table 3**
Comparison of the model accuracy and length of reconstructed $PM_{2.5}$ time series with those in the previous studies over the domain of mainland China.

| Model | Length of $PM_{2.5}$ time series | CV Daily | | Predictive powers Daily | | Monthly | | Reference |
|---|---|---|---|---|---|---|---|---|
| | | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | |
| Two-stage | 2000–2016 | 0.61 | 27.8 | 0.66 | 33.9 | | | Xue et al., 2019 |
| STRF | 2015–2016 | 0.85 | 15.57 | 0.55 | 27.38 | 0.73 | 14.88 | Wei et al., 2019 |
| Two-stage | 2004–2013 | 0.79 | 27.42 | 0.41 | | 0.73 | | Ma et al., 2016 |
| STET | 2000–2018 | 0.89 | 10.33 | | | 0.8 | | Wei et al., 2020 |
| STET | 2000–2018 | 0.86–0.90 | 10–18.4 | | | 0.8 | 11.26 | Wei et al., 2021a |
| LGBM | 1980–2019 | >0.64 | 25.4 | 0.6–0.8 | 16.48 | 0.86 | 9.6 | This study |

with the actual observations. It should be noted that the LGBM model performance in the northwestern China, for example, Xinjiang is not very satisfactory, and this is mainly due to the sparse $PM_{2.5}$ observation stations and interpolation bias in those areas. The application of high-resolution $PM_{2.5}$ concentration retrieved from satellite may overcome this limitation, which need deep investigations in the future studies.

As far as we know, this work is firstly to reconstruct the historical $PM_{2.5}$ concentration in China using LGBM model based on MERRA-2 and ERA-5 meteorological reanalysis data. This study demonstrates that the LGBM is a powerful tool for the correction of the original MERRA-2 $PM_{2.5}$ concentration in China. Additionally, our reconstruction considers only meteorological factors and ignores the emission source changes effects. Further considering the emission source is needed in the future investigation.

## CRediT authorship contribution statement

Jinghui Ma performed formal analysis, writing original draft and conceptualization. Renhe Zhang performed conceptualization, writing-review and editing, and supervision. Jianming Xu performed validation. Zhongqi Yu performed data curation.

## Declaration of competing interest

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

## Acknowledgements

NASA's global modeling and assimilation office is gratefully acknowledged for making the MERRA-2 aerosol reanalysis publicly accessible. This research is supported by the National Natural Science Foundation of China (Grant no. 42005055, 91644223, and 41475040), the Chinese Ministry of Science and Technology (Grant no. 2019YFC0214605), the Natural Science Foundation of Shanghai (Grant no. 19ZR1462100), and the Shanghai Science and Technology Commission (Grant no. 19DZ1205003).

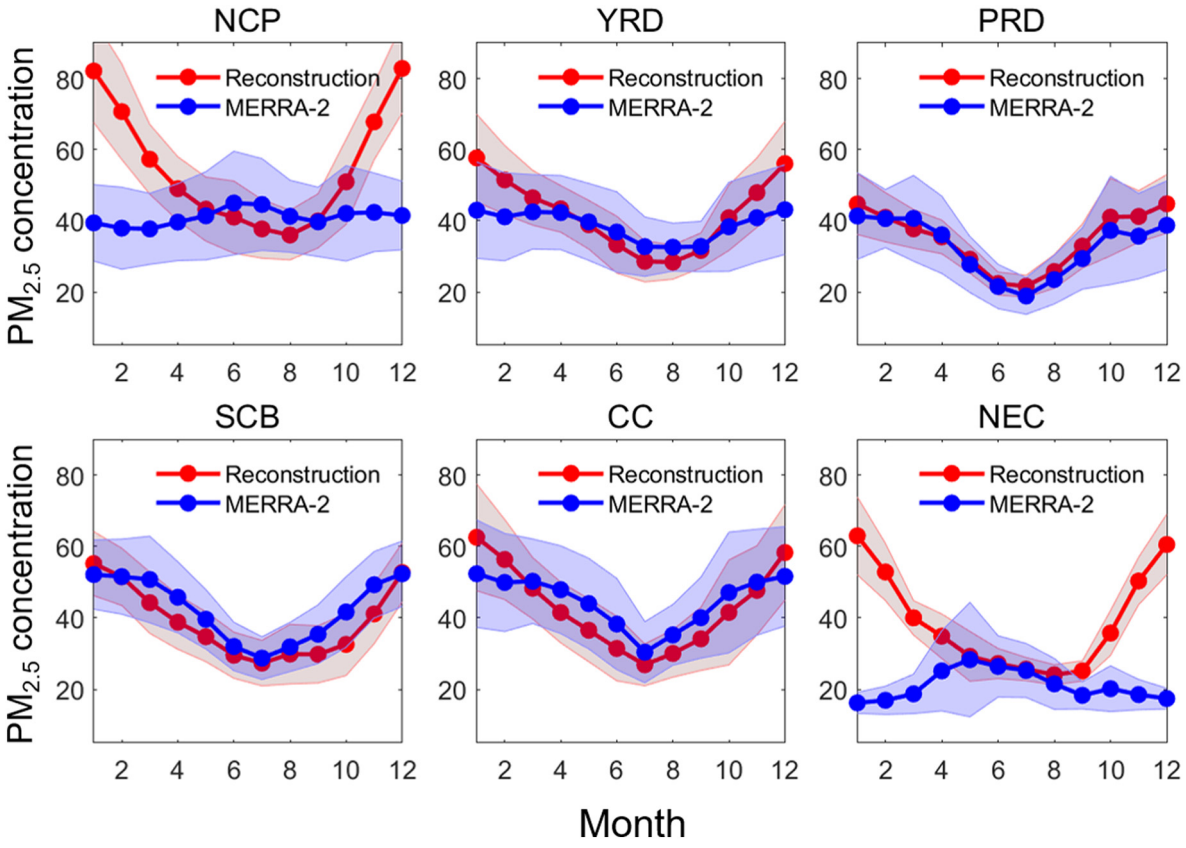



**Fig. 9.** Seasonal variations in monthly mean and standard deviation of original MERRA-2 and reconstructed $PM_{2.5}$ concentrations (unit: $\mu g \cdot m^{-3}$) in 1980–2019 in the six key regions.

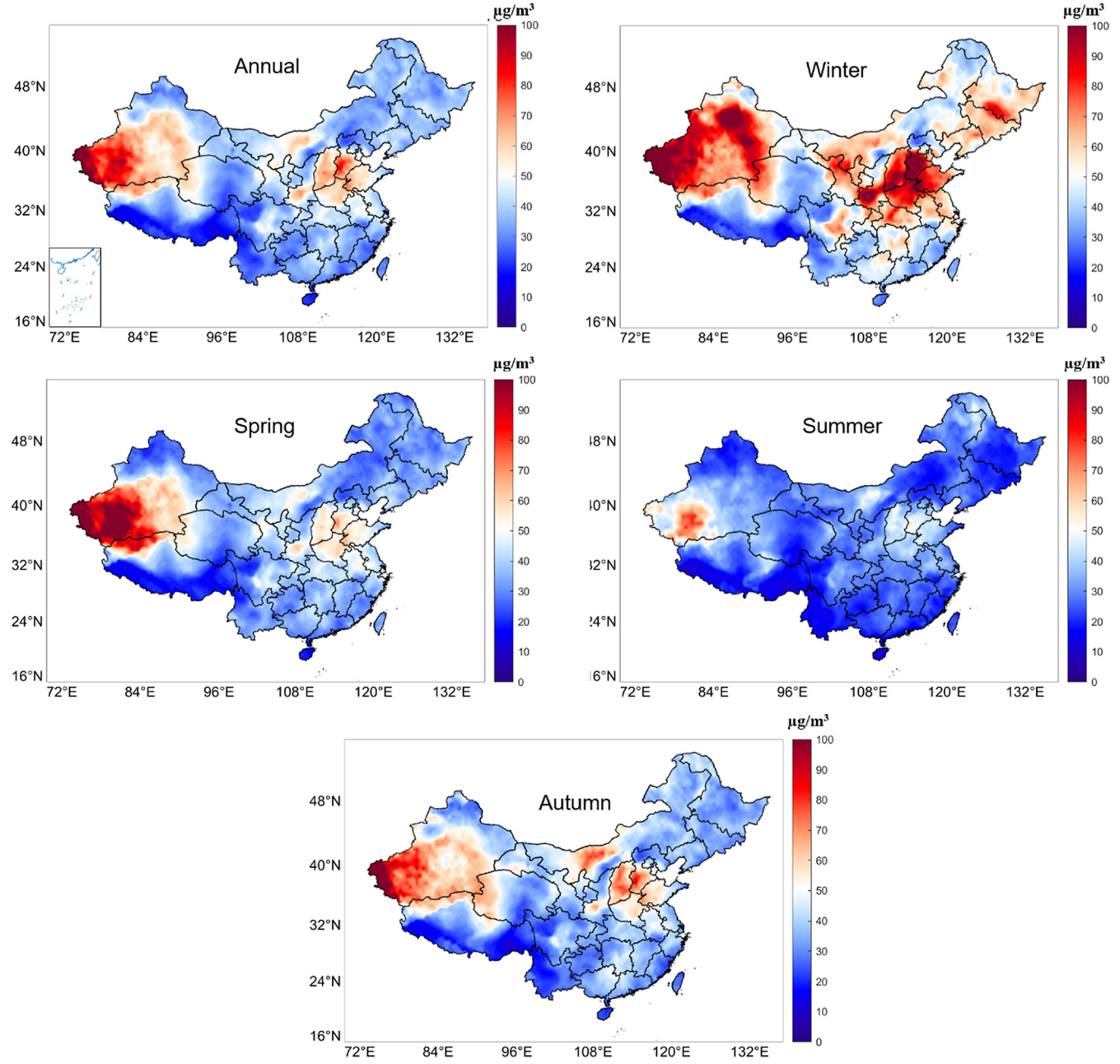



**Fig. 10.** Spatial distributions of reconstructed annual and seasonal $PM_{2.5}$ mass concentrations averaged in 1980–2013 in mainland China.

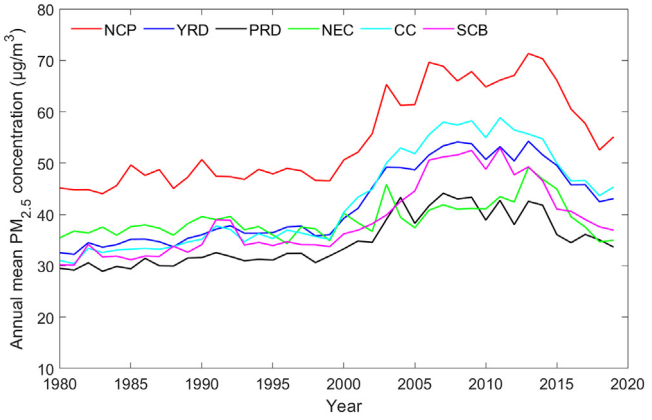



**Fig. 11.** Reconstructed temporal distributions of annual averages of $PM_{2.5}$ across the six key regions of China from 1980 to 2019.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.154363.

## References

An, L., Che, H., Xue, M., Zhang, T., Wang, H., Wang, Y., Zhou, C., Zhao, H., Gui, K., Zheng, Y., Sun, T., Liang, Y., Sun, E., Zhang, H., Zhang, X., 2018. Temporal and spatial variations in sand and dust storm events in East Asia from 2007 to 2016: relationships with surface conditions and climate change. Sci. Total Environ. 633. https://doi.org/10.1016/j.scitotenv.2018.03.068.

Bai, K., Ma, M., Li, K., Li, Z., Guo, J., Chang, N., Tan, Z., Han, D., 2021. LGHAP: a long-term gap-free high-resolution air pollutants concentration dataset derived via tensor flow based multimodal data fusion. Earth Syst. Sci. Data https://doi.org/10.5194/essd-2021-404.

Bao, X., Zhang, F., 2019. How accurate are modern atmospheric re-analyses for the data-sparse Tibetan Plateau region? J. Clim. 32, 7153–7172.

Beers, W., Kleijnen, J., 2003. Kriging for interpolation in random simulation. J. Oper. Res. Soc. 54, 255–262.
Bi, J., Huang, J., Hu, Z., Holben, B., Guo, Z., 2014. Investigating the aerosol optical and radiative characteristics of heavy haze episodes in Beijing during January of 2013. J. Geophys. Res. Atmos. 119 (16), 9884–9900. https://doi.org/10.1002/2014JD021757.
Buchard, V., da Silva, A.M., Randles, C.A., Colarco, P., Ferrare, R., Hair, J., Hostetler, C., Tackett, J., Winker, D., 2016. Evaluation of the surface PM2.5 in version 1 of the NASA MERRA aerosol reanalysis over the United States. Atmos. Environ. 125, 100–111. https://doi.org/10.1016/j.atmose-nv.2015.11.004.
Buchard, V., Randles, C.A., da Silva, A.M., Darmenov, A., Colarco, P.R., Govindaraju, R., Ferrare, R., Hair, J., Beyersdorf, A.J., Ziemba, L.D., Yu, H., 2017. The MERRA-2 aerosol reanalysis, 1980 onward. Part II: evaluation and case studies. J. Clim. 30, 6851–6872. https://doi.org/10.1175/JCLI-D-16-0613.1.
Cai, W., Li, K., Liao, H., Wang, H., Wu, L., 2017. Weather conditions conducive to Beijing severe haze more frequent under climate change. Nat. Clim. Chang. 7, 257–262. https://doi.org/10.1038/nclimate3249.
Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.
Chen, G., Li, S., Knibbs, L., Hamm, N., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M., Guo, Y., 2018. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60.
Chen, Z., Zhang, T., Zhang, R., Zhu, Z., Yang, J., Chen, P., Ou, C., Guo, Y., 2019. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. Atmos. Environ. 202, 180–189.
Dang, R., Liao, H., 2019. Severe winter haze days in the Beijing-Tianjin-Hebei region from 1985 to 2017 and the roles of anthropogenic emissions and meteorological parameters. Atmos. Chem. Phys. Discuss. 5, 1–31. https://doi.org/10.5194/acp-19-10801-2019.
Di, Q., Rowland, S., Koutrakis, P., Schwartz, J., 2017. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. J. Air Waste Manage. Assoc. 67 (1), 39–52.
Diehl, T., Heil, A., Chin, M., Pan, X., Streets, D., Schultz, M., Kinne, S., 2012. Anthropogenic, biomass burning, and volcanic emissions of black carbon, organic carbon, and SO2 from 1980 to 2010 for hindcast model experiments. Atmos. Chem. Phys. Discuss. 12. https://doi.org/10.5194/acpd-12-24895-2012 24 895–24 954.
Ding, A., Huang, X., Nie, W., Chi, X., Xu, Z., Zheng, L., Xu, Z., 2019. Significant Reduction of PM2.5 in Eastern China due to Regional-scale Emission Control: Evidence From SORPES in 2011–2018, pp. 11791–11801.
European Commission, 2011. Commission regulation (EU) No 582/2011 of 25 May 2011 implementing and amending regulation (EC) No 595/2009 of the European Parliament and of the council with respect to emissions from heavy duty vehicles (Euro VI) and amending annexes I and III to directive 2007/46/EC of the European Parliament and of the council. Official J. Eur. Union L 167, 1e168.
Guo, J., Zhang, X., Che, H., Gong, S., An, X., Cao, C., Guang, J., Zhang, H., Wang, Y., Zhang, X., Xue, M., Li, X., 2009. Correlation between PM concentrations and aerosol optical depth in eastern China. Atmos. Environ. 43, 5876–5886.
Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., Shao, J., Yu, T., Tong, B., Li, J., Su, T., Yim, S., Stoffelen, A., Zhai, P., Xu, X., 2021. Investigation of near-global daytime boundary layer height using high-resolution radiosondes: first results and comparison with ERA5, MERRA-2, JRA-55, and NCEP-2 re-analyses. Atmos. Chem. Phys. 21, 17079–17097.
Gui, K., Che, H., Wang, Y., Wang, H., Zhang, L., Zhao, H., Zheng, Y., Sun, T., Zhang, X., 2019. Satellite-derived PM2.5 concentration trends over eastern China from 1998 to 2016: relationships to emissions and meteorological. Environ. Pollut. 247, 1125–1133. https://doi.org/10.1016/j.envpol.2019.01.056.
Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944. https://doi.org/10.1021/acs. est.7b01210.
Huang, J., Liu, J., Chen, B., Nasiri, S., 2015. Detection of anthropogenic dust using CALIPSO lidar measurements. Atmos. Chem. Phys. 15. https://doi.org/10.5194/acp-15-11653-2015 11,653–11,665.
Huang, J., Yin, J., Wang, M., He, Q., Guo, J., Zhang, J., Liang, X., Xie, Y., 2021. Evaluation of five reanalysis products with radiosonde observations over the central Taklimakan Desert during summer. Earth Space Sci. 8, e2021EA001707.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T., 2017. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 3147–3155 2017-Decem.
Li, Q., Zhang, R., Wang, Y., 2016. Interannual variation of the wintertime fog-haze days across central and eastern China and its relation with east asian winter monsoon. Int. J. Climatol. 36, 346–354. https://doi.org/10.1002/joc.4350.
Li, T., Shen, H., Zeng, C., Yuan, Q., Zhang, L., 2017a. Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: methods and assessment. Atmos. Environ. 152, 477–489.
Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017b. Estimating ground-level PM2.5 by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44 (23), 11985–11993.
Li, R., Zhao, Y., Zhou, W., Meng, Y., Zhang, Z., Fu, H., 2020. Developing a novel hybrid model for the estimation of surface 8h ozone (O3) across the remote tibetan plateau during 2005–2018. Atmos. Chem. Phys. 20, 6159–6175. https://doi.org/10.5194/acp-20-6159-2020.
Liao, T., Gui, K., Jiang, W., Wang, S., Wang, B., Zeng, Z., Che, H., Wang, Y., Sun, Y., 2018. Air stagnation and its impact on air quality during winter in Sichuan and Chongqing, southwestern China. Sci. Total Environ. 635, 576–585. https://doi.org/10.1016/j.scitotenv.2018.04.122.
Lin, C., Li, Y., Yuan, Z., Lau, A., Li, C., Fung, J., 2015. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM2.5. Remote Sens. Environ. 156, 117e128.
Liu, J., Weng, F., Li, Z., 2019. Satellite-based PM2.5 estimation directly from reflectance at the top of the atmosphere using a machine learning algorithm. Atmos. Environ. 208, 113–122. https://doi.org/10.1016/j.atmosenv.2019.04.002.
Ma, Z., Hu, X., Sayer, A., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., Liu, Y., 2016. Satellite-based spatiotemporal trends in PM2.5 concentrations: China, 2004–2013. Environ. Health Perspect. 124, 184–192.
Ma, J., Zhang, R., 2020a. Opposite interdecadal variations of wintertime haze occurrence over North China Plain and Yangtze River Delta regions in 1980–2013. Sci. Total Environ. 732, 139240.
Ma, J., Xu, J., Qu, Y., 2020b. Evaluation on the surface PM2.5 concentration over China mainland from NASA's MERRA-2. Atmos. Environ. 237, 117666. https://doi.org/10.1016/j.atmosenv.2020.117666.
Ma, J., Yu, Z., Qu, Y., Cao, Y., 2020c. Application of the XGBoost machine learning method in PM2.5 prediction: a case study of Shanghai. Aerosol Air Qual. Res. 20, 128–138. https://doi:10.4209/aaqr.2019.08.0408.
Mu, M., Zhang, R., 2014. Addressing the issue of fog and haze: a promising perspective from meteorological science and technology. Sci. China Earth Sci. 57, 1–2.
Randles, C.A., da Silva, A.M., Buchard, V., Colarco, P.R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., 2017. The MERRA-2 aerosol reanalysis, 1980 onward. Part I: system description and data assimilation evaluation. J. Clim. 30, 6823–6850. https://doi.org/10.1175/jcli-d-16-0609.1.
Song, Z., Fu, D., Zhang, X., Wu, Y., Xia, X., He, J., Che, H., 2018. Diurnal and seasonal variability of PM2.5 and AOD in North China plain: comparison of MERRA-2 products and ground measurements. Atmos. Environ. 191, 70–78. https://doi.org/10.1016/j.atmosenv.2018.08.012.
Sun, L., Wei, J., Duan, D., Guo, Y., Mi, X., 2016. Impact of land-use and land-cover change on urban air quality in representative cities of China. J. Atmos. Sol. Terr. Phys. 142, 43–54.
Wang, Y., Zhang, X., Sun, J., Zhang, X., Che, H., Li, Y., 2015. Spatial and temporal variations of the concentrations of PM10, PM2.5 and PM1 in China. Atmos. Chem. Phys. 15, 13585–13598.
Wang, P., Guo, H., Hu, J., Harsha, S., Ying, Q., Zhang, H., 2019. Responses of PM2.5 and O3 concentrations to changes of meteorology and emissions in China. Sci. Total Environ. 662, 297–306.
Wang, W., Mao, F., Du, L., Pan, Z., Gong, W., Fang, S., 2017. Deriving hourly PM2.5 concentrations from Himawari-8 AODs over Beijing-Tianjin-Hebei in China. Remote Sens. 9. https://doi.org/10.3390/rs9080858.
Wang, X., Wang, K., Su, L., 2016. Contribution of atmospheric diffusion conditions to the recent improvement in air quality in China. Sci. Rep. 6 (1), 36404.
Wang, X., Zhang, R., 2020a. Effects of atmospheric circulations on the interannual variation in PM2.5 concentrations over the Beijing–Tianjin–Hebei region in 2013–2018. Atmos. Chem. Phys. 20, 7667–7682.
Wang, X., Zhang, R., Tan, Y., Yu, W., 2021. Dominant synoptic patterns associated with the decay process of PM2.5 pollution episodes around Beijing. Atmos. Chem. Phys. 21, 2491–2508.
Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221. https://doi.org/10.1016/j.rse.2019.111221.
Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., Lyapustin, A., Liu, L., Wu, H., Song, Y., 2020. Improved 1 km resolution PM2.5 estimates across China using enhanced space-time extremely randomized trees. Atmos. Chem. Phys. 20 (6), 3273–3289.
Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., Cribb, M., 2021a. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Remote Sens. Environ. 252, 112136. https://doi.org/10.1016/j.rse.2020.112136.
Wei, J., Li, Z., Pinker, R., Wang, J., Sun, L., Xue, W., Li, R., Cribb, M., 2021b. Himawari-8-derived diurnal variations of ground-level PM2.5 pollution across China using the fast space-time light gradient boosting machine (LightGBM). Atmos. Chem. Physi. 21, 7863–7880. https://doi.org/10.5194/acp-21-7863-2021.
Wei, J., Li, Z., Li, K., Dickerson, R., Pinker, R., Wang, J., Liu, X., Sun, L., Xue, W., Cribb, M., 2022. Full-coverage mapping and spatiotemporal variations of ground-level ozone (O3) pollution from 2013 to 2020 across China. Remote Sens. Environ. 270, 112775. https://doi.org/10.1016/j.rse.2021.112775.
Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. 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. 201, 345–357.
Zeng, Z., Wang, Z., Gui, K., Yan, X., Gao, M., Luo, M., Geng, H., Liao, T., Li, X., An, J., Liu, H., He, C., Ning, G., Yang, Y., 2020. Daily global solar radiation in China estimated from high-density meteorological observations: a random forest model framework. Earth Sp. Sci. https://doi.org/10.1029/2019ea001058.
Zhang, X., Wang, Y., Zhang, X., Guo, W., Gong, S., 2008. Carbonaceous aerosol composition over various regions of China during 2006. J. Geophys. Res. 113, D14111. https://doi.org/10.1029/2007JD009525.
Zhang, R., Li, Q., Zhang, R., 2014. Meteorological conditions for the persistent severe fog and haze event over eastern China in January 2013. Sci. China Earth Sci. 57, 26–35.
Zhang, R., 2017. Warming boosts air pollution. Nat. Clim. Chang. 7, 238–239.
Zhang, X., Xu, X., Ding, Y., Liu, Y., Zhang, H., Wang, Y., Zhong, J., 2019. The impact of meteorological changes from 2013 to 2017 on PM2.5 mass concentration in key regions in China. Sci. China Earth Sci. 62, 1885–1902.
Zhang, Y., Li, Z., Chang, W., Zhang, Y., Leeuw, G., James, J., 2020. Satellite observations of PM2.5 changes and driving factors based forecasting over China 2000–2025. Remote Sens. 12, 2518. https://doi.org/10.3390/rs12162518.
Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., Zhang, W., 2021. Robust prediction of hourly PM2.5 from meteorological data using LightGBM, natl. Sci. Rev. 8, nwaa307. https://doi.org/10.1093/nsr/nwaa307.