Environment International 134 (2020) 105297

Contents lists available at ScienceDirect

Environment International

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

Estimating the daily $PM_{2.5}$ concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01° × 0.01° spatial resolution

Chen Zhao[a,b,1], Qing Wang[a,1], Jie Ban[a,1], Zhaorong Liu[b,*], Yayi Zhang[a], Runmei Ma[a], Shenshen Li[c], Tiantian Li[a,*]

[a] National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100021, China
[b] College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China
[c] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Handling Editor: Da Chen

Keywords:
$PM_{2.5}$ estimation
Human exposure
High spatiotemporal resolution
Machine learning

ABSTRACT

High spatiotemporal resolution fine particulate matter ($PM_{2.5}$) simulations can provide important exposure data for the assessment of long-term and short-term health effects. Satellite-based aerosol optical depth (AOD) data, meteorological data, and topographic data have become key variables for $PM_{2.5}$ estimation. In this study, a random forest model was developed and used to estimate the highest resolution (0.01° × 0.01°) daily $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region. Our model had a suitable performance (cv-R^2 = 0.83 and test-R^2 = 0.86). The regional test-R^2 value in southern Beijing-Tianjin-Hebei was higher than that in northern Beijing-Tianjin-Hebei. The model performance was excellent at medium to high $PM_{2.5}$ concentrations. Our study considered meteorological lag effects and found that the boundary layer height of the one-day lag had the most important contribution to the model. AOD and elevation factors were also important factors in the modeling process. High spatiotemporal resolution $PM_{2.5}$ concentrations in 2010–2016 were estimated using a random forest model, which was based on $PM_{2.5}$ measurements from 2013 to 2016.

1. Introduction

In recent years, the long-term and short-term health effects of fine particulate matter ($PM_{2.5}$) have increasingly been the focus of environmental epidemiological studies (Lu et al., 2017; Hu et al., 2014; Kheirbek et al., 2013). Studies have confirmed that $PM_{2.5}$ exposure is positively correlated with the morbidity and mortality rates of various diseases, such as respiratory diseases, cardiovascular diseases, and cerebrovascular and lung cancer (Hu et al., 2014; Fang et al., 2013). The exposure concentrations are mostly directly monitored by $PM_{2.5}$ monitoring stations, but the monitoring sites have low spatial coverage (Yi et al., 2018), and the differences in the $PM_{2.5}$ concentration cannot be captured at a fine scale. The full coverage of high-spatial-resolution $PM_{2.5}$ estimates is necessary for the refinement of exposure data used in environmental epidemiology. $PM_{2.5}$ estimates with high spatiotemporal resolution can provide basic data on the exposure-response relationship between human health conditions and $PM_{2.5}$ concentrations. As a typical polluted area in China, the Beijing-Tianjin-Hebei region has a high population density and high emission of pollutants (Zhao et al., 2011). The exposure of the population has been at a high level for a long time.

There are three challenges to obtain high spatiotemporal resolution $PM_{2.5}$ simulations. First, ground monitoring stations are sparse, which weakens the $PM_{2.5}$ simulation effect. With the development of satellite remote sensing technology, aerosol optical depth (AOD) measurements with high spatial and temporal resolutions (e.g., 10-km medium resolution imaging spectroradiometer (MODIS) AOD measurements) (Ma et al., 2016; Yang and Hu, 2018) have gradually become some of the vital basic data in $PM_{2.5}$ concentration simulations. To the best of our knowledge, few studies have used 3-km MODIS AOD data to estimate $PM_{2.5}$ concentrations at high spatiotemporal resolutions. In addition, important variables related to $PM_{2.5}$ concentrations should be considered to improve the simulation accuracy, such as meteorological variables, topographical factors, and land use factors (Zheng et al., 2016; Hu et al., 2017; Cattani et al., 2017). Second, there is a need to estimate $PM_{2.5}$ concentrations at high spatial resolutions for environmental epidemiological studies. The determination of the differentiations in human $PM_{2.5}$ exposure levels is important for determining the long-term health effects of $PM_{2.5}$ pollution. The current spatial resolution of daily $PM_{2.5}$ simulations in the Beijing-Tianjin-Hebei region is approximately 10 km × 10 km (Yang and Hu, 2018; Wang et al., 2016).

* Corresponding authors.
E-mail addresses: zrliu@pku.edu.cn (Z. Liu), litiantian@nieh.chinacdc.cn (T. Li).
[1] These authors contributed equally.

https://doi.org/10.1016/j.envint.2019.105297
Received 7 May 2019; Received in revised form 28 October 2019; Accepted 29 October 2019
Available online 27 November 2019

Over a smaller scale range, such as Beijing, there have been studies with a spatial resolution of 1 km × 1 km (Liang et al., 2018). However, to the best of our knowledge, $PM_{2.5}$ simulations with a 1 km × 1 km spatial resolution have not been carried out in the Beijing-Tianjin-Hebei region. Third, long-term $PM_{2.5}$ simulations lack historical $PM_{2.5}$ monitoring data. A $PM_{2.5}$ monitoring network was built in China that covered the whole country at the end of 2012 (Ma et al., 2016). Long-term exposure is usually simulated with $PM_{2.5}$ measurements over consecutive years. For example, Ma et al. (Ma et al., 2016) used the $PM_{2.5}$ measurements from 2013 to simulate the $PM_{2.5}$ concentrations from 2004 to 2012. The aforementioned method might have led to interannual uncertainties when simulating long-term historical $PM_{2.5}$ concentrations using only short-term $PM_{2.5}$ measurements.

Many statistical models have been developed and applied to assess the exposure to $PM_{2.5}$ concentrations. At present, the widely used and well-performing models include land use regression models (Shi et al., 2016; Hu et al., 2016; Eeftens et al., 2012), linear mixed-effects models (Zheng et al., 2016; Ma et al., 2016), geographically weighted regression models (van Donkelaar et al., 2015; Zhang et al., 2016; You et al., 2016), generalized additive models (Wang et al., 2016; Yanosky et al., 2014; Kloog et al., 2014), remote sensing formulas (Lv et al., 2017; Lv et al., 2016), geographically weighted gradient boosting machine learning models (Zhan et al., 2017), random forest models (Chen et al., 2018; Hu et al., 2017) and neural network models (Li et al., 2017; Di et al., 2016). Previous studies have shown that random forest and neural network models exhibit excellent performance in $PM_{2.5}$ simulations (Chen et al., 2018; Li et al., 2017). In contrast to the neural network model, the random forest model avoids complex structures (Hu et al., 2017; Di et al., 2016) and consumes fewer computational resources. The importance of variables is provided by the random forest model to guide the screening of model variables. Building a statistical model that is based on annual $PM_{2.5}$ monitoring data is a universal approach, which can reduce annual variations and ensure accurate simulations.

In our study, we estimated the $PM_{2.5}$ concentrations at high spatiotemporal resolution in the Beijing-Tianjin-Hebei region, which is a typical polluted area in China, and provided accurate exposure data for epidemiological studies on the long-term and short-term health effects of $PM_{2.5}$ pollution. By constructing a high spatiotemporal resolution (0.01° × 0.01°) random forest $PM_{2.5}$ model in the Beijing-Tianjin-Hebei region, we aimed to (1) consider the lag effects of meteorological conditions to provide a reference for the important factors in model construction for future research; (2) explore the method for simulation of historical $PM_{2.5}$ measurements (2010–2012) and the refinement of existing measurements (2013–2016); (3) estimate the daily $PM_{2.5}$ concentrations (2010–2016) in the Beijing-Tianjin-Hebei region at high spatiotemporal resolution.

2. Materials and methods

2.1. Study area

The main research area of the Beijing-Tianjin-Hebei region (113.45°E-119.85°E and 36.03°N-42.62°N) and the extended area (112.95°E-120.35°E and 35.53°N-43.12°N) were selected to prevent boundary effects, and the study area included 43 prefecture-level cities in 8 Chinese provinces and municipalities (Fig. 1). The Beijing-Tianjin-Hebei region is the largest economic circle in northern China with a high population density. High particulate emissions, special topographic conditions and meteorological conditions significantly contribute to the formation of high $PM_{2.5}$ concentrations. The study period was from January 1st, 2010, to December 31st, 2016, which included 2557 days in total. Using the model constructed based on the $PM_{2.5}$ measurements from 2013 to 2016, the daily $PM_{2.5}$ concentrations from 2010 to 2016 were estimated.

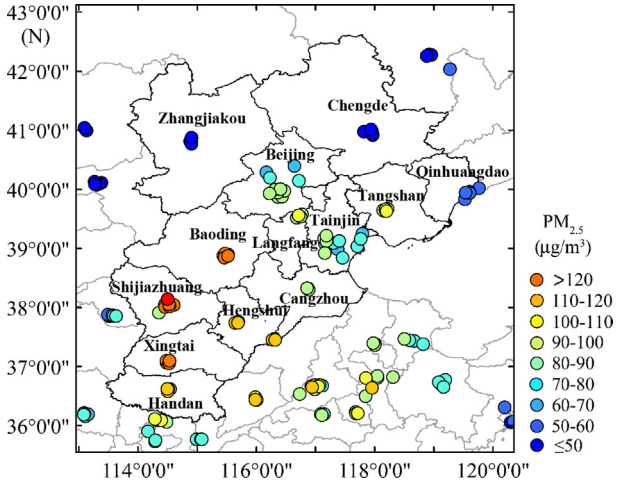

Fig. 1. Study area map showing 156 regional monitoring sites and mean ground-level-measured $PM_{2.5}$ (μg/m^3) concentrations from 2013 to 2016.

2.2. Datasets

2.2.1. Ground-based $PM_{2.5}$ measurements

Daily ground $PM_{2.5}$ measurements were taken from January 2013 to December 2016 from the China Environmental Monitoring Center (http://www.cnemc.cn/). There were a total of 156 national air quality monitoring sites (Fig. 1) in the study area. Tapered element oscillating microbalance technology was used to measure the $PM_{2.5}$ concentration at all sites. No traffic sites were included in the model. The mean value of the $PM_{2.5}$ measurements was calculated for all sites (Fig. 1).

2.2.2. AOD data

MODIS is equipped on the Earth Observing System (EOS) series of satellites from the United States and is used for atmospheric AOD inversion. MODIS level 2 aerosol data represent one of the best AOD products for near real-time aerosol data assimilation (Remer et al., 2013). Aqua MODIS level 2 data (C6) (with a spatial resolution of 3 km × 3 km) of the study area were obtained (https://ladsweb.modaps.eosdis.nasa.gov/) for the period from January 1, 2010, to December 31, 2016.

Missing data constitute the inadequacies of the current AOD data. The lack of data is mainly caused by high degrees of ground reflection and cloud influences. Our previous study constructed random forest models to simulate missing values in the current study area by establishing a nonlinear relationship between the AOD and multivariate analysis results (Zhao et al., 2019). We simulated the missing AOD values in the current study.

2.2.3. Meteorological fields

The meteorological data during the study period (7 years) were obtained from the ERA-Interim reanalysis data of the European Center for Medium-Range Weather Forecasts (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/). We considered the impact of meteorological hysteresis on the model, and we included two days of meteorological lag data in the model. Forty-five types of meteorological data were adopted in our model (Table S1). These meteorological data used daily surface data with a spatial and temporal resolutions of 0.125° × 0.125° and 6/12 h, respectively.

2.2.4. Land use variables

Annual land use data from 2010 to 2016 at a spatial resolution of 300 m were downloaded from the European Space Agency Climate Change Initiative (http://maps.elie.ucl.ac.be/CCI/viewer/) website. The data were collected using 300-m medium resolution imaging spectrometer (MERIS), full resolution (FR) and reduced resolution (RR)

products, as well as the EOS of the French Space Research Center, the advanced very high resolution radiometer (AVHRR) sensor mounted on the National Oceanic and Atmospheric Administration (NOAA) meteorological satellite, and the Project for On-Board Autonomy-Vegetation (PROBA-V). Three types of land use data, namely, natural vegetation coverage, urban coverage, and farmland coverage, were extracted.

2.2.5. Elevation data

The elevation data (with a spatial resolution of 1 km × 1 km) for 2010 came from the Resource and Environmental Science Data Center (RESDC) of the Chinese Academy of Sciences (http://www.resdc.cn).

2.2.6. Population

The national population data for 2010 were collected from the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn), and were distributed on a 1 km × 1 km grid. Considering the geographical differences of the population-natural elements that were generated by spatial interpolation of the 1 km × 1 km grid data, the data of this kilometer grid distribution were based on the country's demographic data.

2.2.7. Road network

The road network data for 2016 were published by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn). The latest published data have taken advantage of the information in the latest navigation maps. Highways, railways, national highways, provincial highways, county roads, country roads, other vehicular roads and pedestrian roads have been included as road elements. In our study, national highways, highways, provincial highways, railways, county roads and country roads were adopted.

2.2.8. Chemical transport model of $PM_{2.5}$ output

Based on data availability differences, the Community Multiscale Air Quality (CMAQ) model was adopted to present the $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region during 2012–2015 at various resolutions: 4 km (January, April, July and October 2012; April, August and October 2013; December 2014; and November and December 2015) or 12 km (January 2013 and January, March, July, October and November 2014) spatial resolution (Hu et al., 2016). The model used the CB05-AE6 or two-dimensional volatility basis set (2D-VBS) mechanism. The CB05-AE6 mechanism is the updated CB05 algorithm and utilizes the chemical compositions of liquids and aerosols. Simulated meteorological fields (the Weather Research Forecast (WRF) 3.7 model) and emission inventory data were used in the CMAQ model.

2.2.9. Regional and seasonal dummy

Previous studies noted that the concentrations of pollutants in $PM_{2.5}$ concentrations were affected by geographical factors and impacts of emissions from different regions. A difference in the $PM_{2.5}$ concentrations between provinces and municipalities in the Beijing-Tianjin-Hebei region was previously noted. The seasonal $PM_{2.5}$ concentration was affected by atmospheric discharge conditions, such as pollution source discharge, precipitation, and wind speed. To account for the latter, we added regional and seasonal dummy variables. The regional variables were divided into 8 regional variables by dividing the provinces and municipalities: Beijing, Tianjin, Hebei, Inner Mongolia, Liaoning, Shanxi, Henan, and Shandong. There were four seasonal variables for the spring, summer, autumn and winter.

2.3. Data processing

The entire study area was created to extend the Beijing-Tianjin-Hebei boundary range by 0.5° (Fig. 1) (East longitude 112.95°-120.35° and North latitude 35.55°-43.12°) to reduce the boundary effects of the $PM_{2.5}$ simulations. A standard grid with a spatial resolution of 0.01° × 0.01° was created. All data were resampled to standard network grids by bilinear interpolation. Nearby 0.03° × 0.03° standard grids were used to extract and calculate three types of land use coverages (natural vegetation coverage, urban coverage and farmland coverage). The road length in each normalized grid was extracted by vector segmentation of the data. The $PM_{2.5}$ concentration output by the chemical transport model, which was calculated as the monthly average, was applied to the corresponding quarter. The regional and seasonal dummy variables were encoded using one hot spot. All data were integrated into a standard grid. According to the latitude and longitude values of the points, data were extracted as training datasets. Data normalization was applied during data preprocessing, which aimed to eliminate dimensional influences. The normalization can be expressed in general as follows:

$$X_{newij} = (X_{ij} - X_{mean})/X_{std} \quad (1)$$

where X_{newij} is the value of data normalization on day $i$in grid cell j; X_{ij} represents the raw data of the variables extracted from the standardized grids; and X_{mean} and X_{std} are the mean and standard deviation of X_{ij}, respectively. ArcGIS and batch processing in Python 2.7.13 were utilized for data processing and figure drawing.

2.4. Established models

The random forest model consisted of multiple decision trees that were generated by the bagging ensemble method (which was enhanced by combining relatively weak models). The decision tree used a tree branch structure to achieve classification and built the structure through decision data based on the best separation point. The final random forest model was expressed as follows:

$$PM_{2.5ij} \sim AOD_{ij} + METE_{ij} + lag1\_METE_{ij} + lag2\_METE_{ij} + POP_j + ELEV_j + ROAD_j + LD_{ij} + CH\_PM_{2.5ij} + SEAS_{ij} + PROV_j \quad (2)$$

where $PM_{2.5ij}$ is the $PM_{2.5}$ concentration on day i in grid cell j; AOD, $METE$, POP, $ELEV$, $ROAD$, and LD are AOD, meteorological variable, population, elevation, length of the road, and land use coverage in grid cell j, respectively; $lag1\_METE$ and $lag2\_METE$ are the meteorological variables of the one-day lag and two-day lag; $CH\_PM_{2.5}$ is the monthly average that matches the corresponding quarter; and $SEAS$ and $PROV$ are the season of year and province of the study area, respectively. During model parameter adjustments, it is often necessary to adjust two parameters (the maximum depth (the number of predictors sampled for the splitting process at each node) and $n_{estimators}$ (the number of trees grown)) in the random forest application in Python. We used the cart decision tree in our study to construct the random forest model. The error rate was calculated by employing predictions of out-of-bag samples. By manual testing, we set the maximum depth as 55 and $n_{estimators}$ as 200 to achieve a high prediction accuracy in the experiments.

The $PM_{2.5}$ measurements were used as dependent variables to construct the random forest model, and the other data were used as independent variables. We employed all of the training datasets in the model simulations to find the generalization of the model and simulated the historical data in the study area. The estimated historical daily $PM_{2.5}$ concentrations (2010–2012) were calculated by employing the built random forest model based on 2013–2016 data because China had not built a $PM_{2.5}$ monitoring network before 2013. The $PM_{2.5}$ concentrations in 2013–2016 were estimated using the same random forest model based on 2013–2016 data. All modeling work was performed in Python 2.7.13 using the scikit-learn package.

2.5. Validation

A 10-fold cross-validation (cv) and test set evaluation were included in the verification method. A 10-fold cv process was performed by

randomly dividing the training dataset into 10 subsets, which were evaluated a total of ten times. Each time, 9 subsets were used as training datasets to verify the data and were then evaluated with the remaining subset. The final result was the mean of the 10 evaluated results. Test set evaluation was used as the final verification method to increase the reliability of the results. In our study, all modeling data were divided according to a 9:1 ratio with regard to the training and test datasets. The model evaluation indicators included the R^2 value of the 10-fold cv process (cv-R^2), R^2 value of the test set evaluation (test-R^2), mean prediction error of the test set evaluation (test-MPE) and root mean square prediction error of the test set evaluation (test-RMSPE).

We also employed other validation methods to verify the stability of the model: (1) Monitoring data from 90% of the stations that were randomly selected were used as the training set, while the monitoring data from the remaining stations were used as the testing set. (2) Monitoring data for ten days in January, April, July and October of each year were randomly selected as the testing set, for which 160 days in total were selected. Monitoring data for the remaining days were used as the training set. These two methods focused on different random divisions of the data set, and the entire modeling process was the same as that used in our main analysis. Furthermore, considering that many models currently have poor capture capabilities for high values, we also used our main model to calculate the R^2 for the high-concentration sets with concentrations above 75 μg/m^3 and 115 μg/m^3, which are considered as the lower bounds of the China air quality standard used for defining light and moderate $PM_{2.5}$ pollution levels, respectively. Furthermore, the 95th and 98th percentiles of the monitoring data were also used as standards. Model performance of the low values was also considered by calculating R^2 for low-concentration sets with the same standards.

3. Results

A descriptive statistical analysis of all variables in the $PM_{2.5}$ model construction is shown in Table S2 with a total of 178,507 modeling data points. We also visualized some of the parameter data (Fig. S1). The mean $PM_{2.5}$ measurement and standard deviation were 79.23 μg/m^3 and 68.66 μg/m^3, respectively. The measured concentration of $PM_{2.5}$ in the Beijing-Tianjin-Hebei region was much higher than that in the United States, which was 7.5 μg/m^3 in 2012 (Di et al., 2016). The mean and standard deviation values of the full coverages of the high spatial and temporal resolution daily AOD estimations were 0.63 and 0.29, respectively (Table S2).

The simulation results of the $PM_{2.5}$ estimations and the $PM_{2.5}$ measurements at the daily level are shown in Fig. 2, with a cv-R^2 of 0.83 and a test-R^2 of 0.86. Fig. 3 shows the model performance between the different years and seasons at the daily level, which indicated the stability of the model on a temporal scale. The test-R^2 had a maximum of 0.88 in 2016, and the lowest value was 0.84 in 2013 and 2014. The model's test-R^2 value ranged from 0.79 to 0.88 in the four seasons. The $PM_{2.5}$ observations and the simulated $PM_{2.5}$ values were lower in summer, which is when the model exhibited the worst performance. The model simulation effect was the best in autumn, and the model simulation performances were second best in spring and winter.

The simulation results of the model in the different provinces, municipalities and prefecture-level cities were evaluated, and the different effects of the model performances were revealed. Table 1 shows that the model's prediction results in Beijing (test-R^2 = 0.91) were higher than those in Tianjin (test-R^2 = 0.89) and Hebei (test-R^2 = 0.89) (Fig. S4). Our model showed stability at different spatial scales. The root mean square error (RMSE) and mean absolute error (MAE) in Tianjin were 19.5 and 12.4, respectively, which were lower than those in Beijing (test-RMSE = 20.81 and test-MAE = 12.69) and Hebei (test-RMSE = 24.05 and test-MAE = 14.36). The model simulation results for the prefecture-level cities in Hebei Province were the best in Tangshan, Langfang and Xingtai, and the test-R^2 values were higher

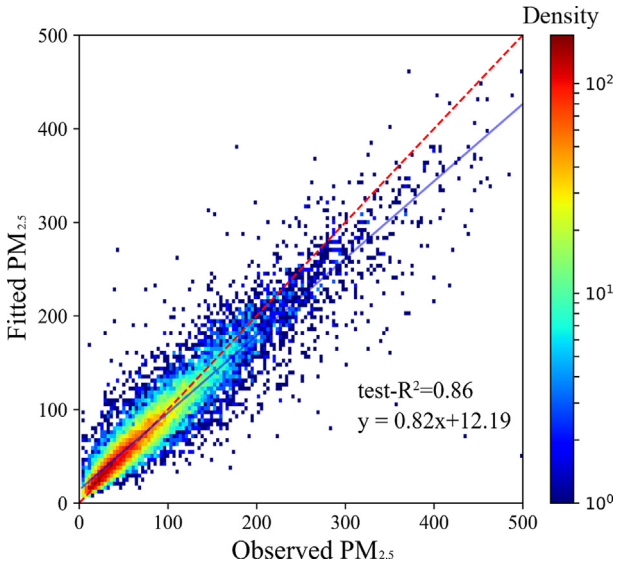

Fig. 2. Density scatter plots of the model fitting and test set validation results at the daily level (n = 17851).

than 0.9. Qinhuangdao, Chengde and Zhangjiakou are located in the northern part of the study area, and the model performed poorly in this area. The model also showed the lowest simulation results (test-R^2 = 0.75) in Qinhuangdao.

The statistical analysis results of the annual averages of the model simulation are shown in Table S4. The model simulation results of the $PM_{2.5}$ concentration had high spatial and temporal resolutions (0.01° × 0.01°, daily). The annual $PM_{2.5}$ concentrations of the historical data obtained by the model for 2010–2012 were 63.51 μg/m^3, 66.63 μg/m^3, and 70.52 μg/m^3, respectively. The annual $PM_{2.5}$ estimations in the Beijing-Tianjin-Hebei region from 2013 to 2016 were 76.94 μg/m^3, 75.73 μg/m^3, 67.13 μg/m^3, and 62.24 μg/m^3, respectively. There was a similar distribution trend between the simulated $PM_{2.5}$ concentration (Fig. 4) and the observed distribution (Fig. S2). Fig. 4 shows that in the southern part of the study area, the $PM_{2.5}$ estimations were higher than those in the northern part of the study area, specifically in Shijiazhuang, Xingtai and Handan. The $PM_{2.5}$ estimations showed the highest concentration trend near the Taihang Mountains. The annual $PM_{2.5}$ concentration in Beijing was lower than that in the southern region but was significantly higher than that in the northern region, namely, in Zhangjiakou and Chengde, which showed lower $PM_{2.5}$ concentrations.

Fig. 5 shows the seasonal distribution of the mean of the estimated $PM_{2.5}$ concentrations, which revealed clear seasonal changes. In the winter, the southern part of the Beijing-Tianjin-Hebei region showed a high concentration of $PM_{2.5}$, and in the summer, this region exhibited a low concentration. The winter $PM_{2.5}$ estimations for the 2010–2016 winters were 79.3 μg/m^3, 98.11 μg/m^3, 103.77 μg/m^3, 108.9 μg/m^3, 83.53 μg/m^3, 84.83 μg/m^3, and 113.72 μg/m^3, respectively (Table S4). The winter of 2016 included only December. The seasonal mean of the $PM_{2.5}$ measurements in the southern part was higher than 100 μg/m^3 in the winters of 2013–2016, and the pollution was the most serious in the winters of 2013 and 2016 (Table S5). The $PM_{2.5}$ estimations exhibited similar results (Table S6). Near the Taihang Mountains, the $PM_{2.5}$ concentration was relatively high in the different seasons. The simulated values showed that the $PM_{2.5}$ concentrations in the different seasons were lower in northern Beijing than those in southern Beijing.

The stability of the model was verified through various validation results. The performance for the model that randomly selected the sites for nine-one segmentation was good, with the test-R^2 reaching 0.89 and the RMSE reaching 14.04 (Fig. S5). For the validation method that randomly selected days from several months, the test-R^2 was 0.56 and

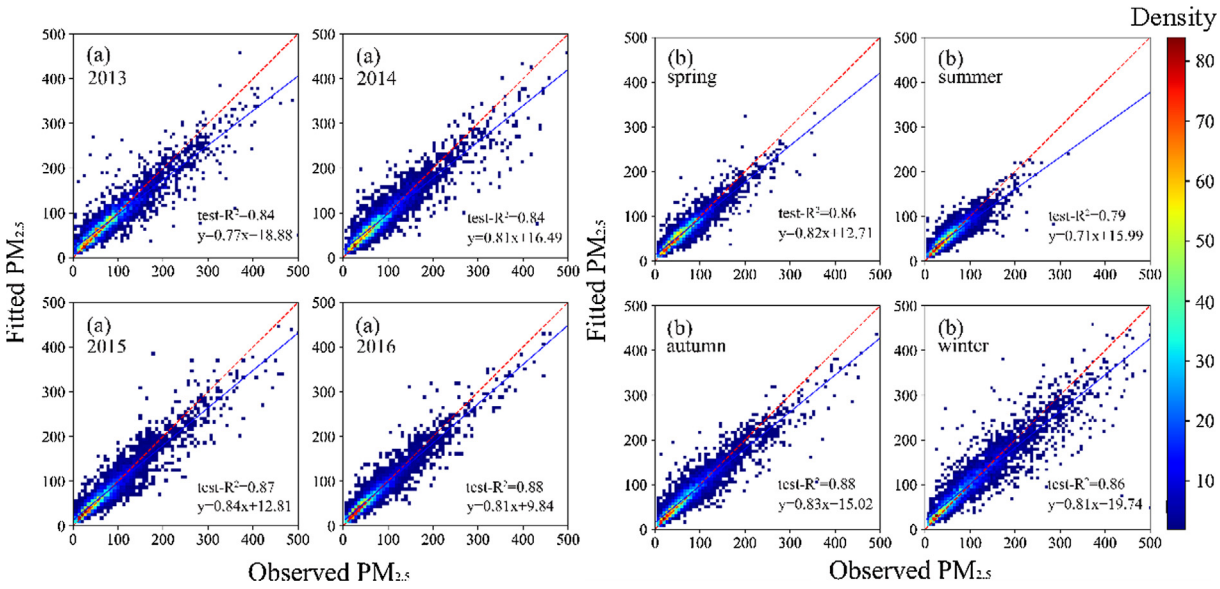

Fig. 3. Evaluation of the estimation performance of the random forest model by using test set validation at the daily level. (a) Annual performance. (b) Seasonal performance.

Table 1
Model performances of the overall region, different provinces, and prefecture-level cities.

Province region	Test-R^2	RMSE	MAE	Prefecture-level city	Test-R^2	RMSE	MAE
Total	0.86	23.48	13.65				
Beijing	0.91	20.81	12.69	–	–	–	–
Tianjin	0.89	19.5	12.4	–	–	–	–
Hebei	0.86	24.05	14.36	Baoding	0.88	30.39	17.8
				Cangzhou	0.87	21.61	13.5
				Chengde	0.82	14.31	9.47
				Handan	0.88	29.8	17.81
				Hengshui	0.89	23.85	14.11
				Langfang	0.93	19.61	12.67
				Qinhuangdao	0.75	21.88	15.37
				Shijiazhuang	0.88	28.86	18.31
				Tangshan	0.94	16.72	11.38
				Xingtai	0.92	26.27	16.59
				Zhangjiakou	0.77	13.96	8.74

the RMSE was 45.91 (Fig. S6). The model's ability to capture high values was good. For high concentrations that exceeded 75 μg/m^3, the test R^2 was 0.83 and the RMSE was 30.08; for concentrations higher than 115 μg/m^3, the test R^2 was 0.76, and the RMSE was 37.70 (Fig. S7). The test R^2 values for concentration sets lower than 75 μg/m^3 or 115 μg/m^3 were 0.41 and 0.70, and the RMSEs were 14.00 and 15.30, respectively (Fig. S7). More results about the model performance are shown in the Supplementary Information.

4. Discussion

A random forest model was developed and used to estimate the $PM_{2.5}$ concentrations at high spatiotemporal resolution in the Beijing-Tianjin-Hebei region. To the best of our knowledge, our research has the highest spatial resolution (0.01° × 0.01°) at the daily level in the study area. A suitable performance was achieved by the multivariate random forest model. The overall simulation test-R^2 value reached 0.86, and the cv-R^2 was 0.83. The test-R^2 value of the model showed a trend of increasing each year, and the seasons and regions with high

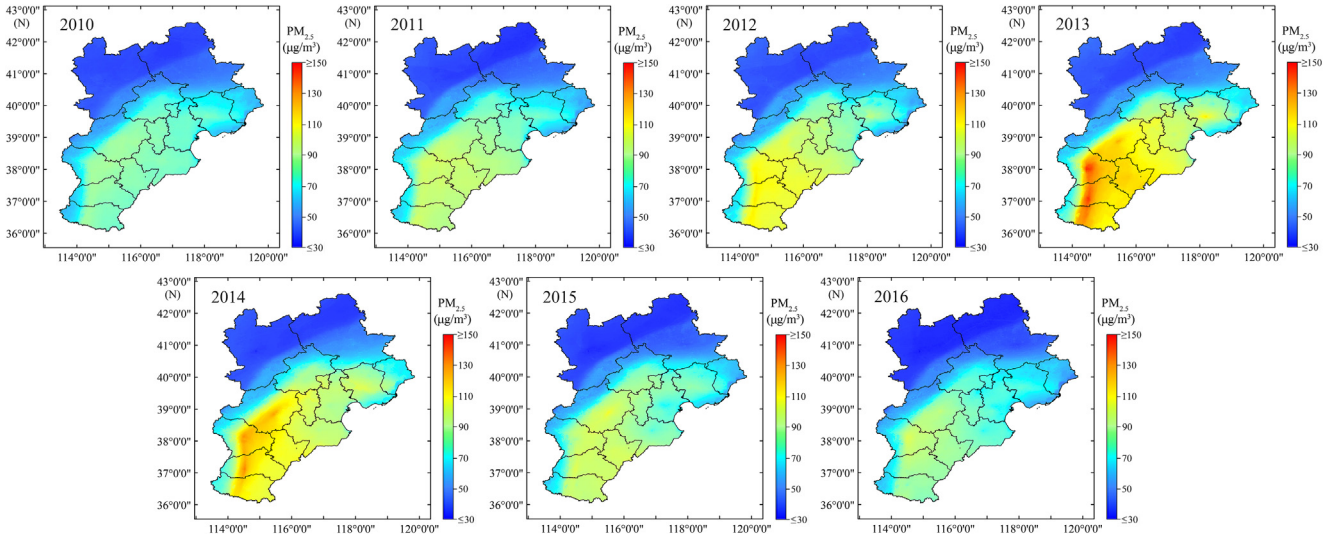

Fig. 4. The model estimates of the average annual $PM_{2.5}$ concentrations in the study area from 2010 to 2016. The red color indicates high values, and the blue color indicates low values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

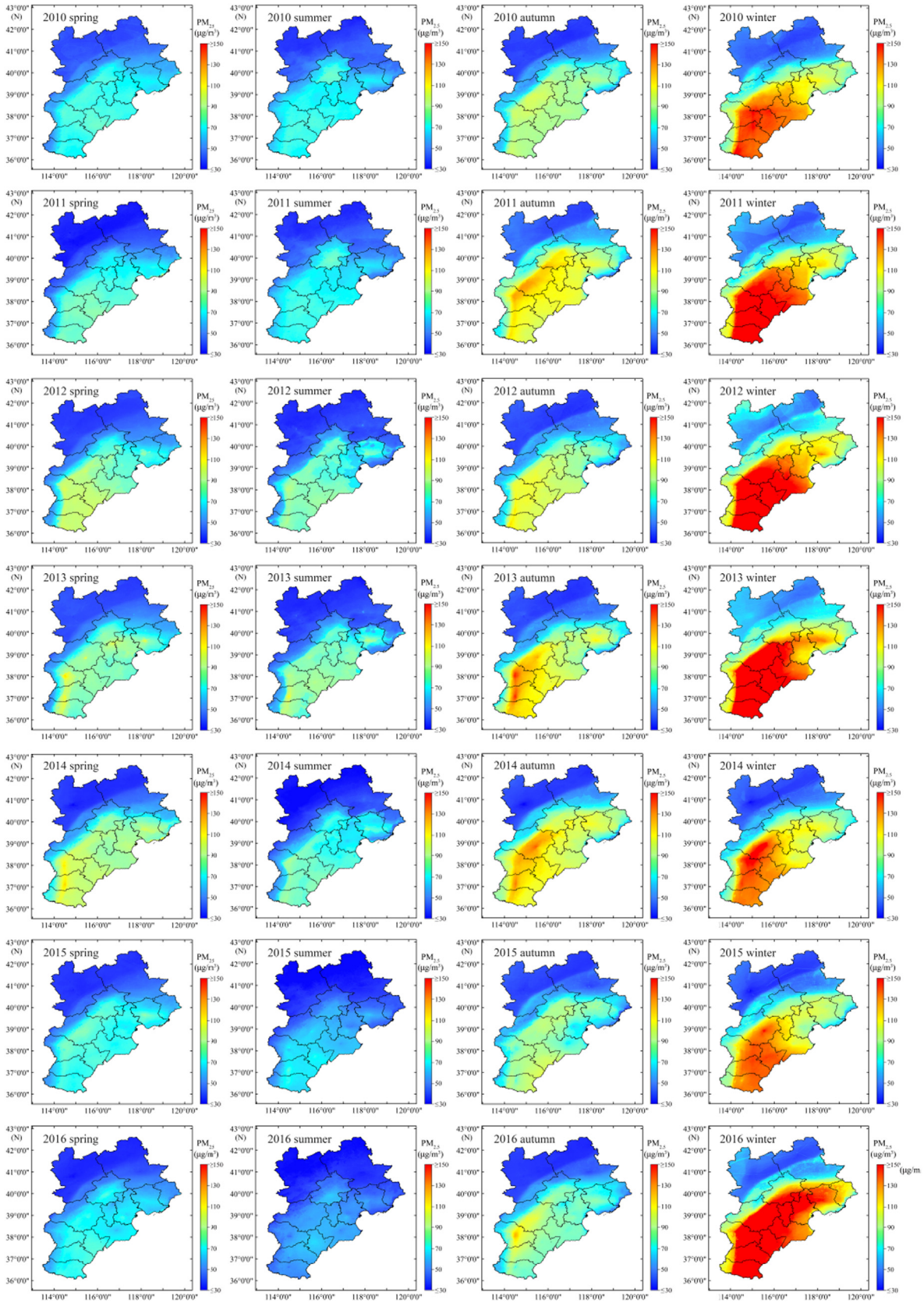

Fig. 5. The model estimates of the average seasonal $PM_{2.5}$ concentrations of the different years in the study area from 2010 to 2016. (Spring was defined as March to May; summer was defined as June to August; autumn was defined as September to November; and winter was defined as December to February of the following year. In particular, the winter of 2016 included only December. Red indicates high values, and blue indicates low values.)

$PM_{2.5}$ concentrations showed better model performances. The addition of meteorological lag data significantly improved the model performance. The data on the one-day lag of the boundary layer height made the largest contribution to the model, and the feature importance was 21.12%. Our research showed the crucial impact of the meteorological lag on the model, which can provide a reference for feature selection for more accurate $PM_{2.5}$ estimates in future research. We constructed a generic model using extended $PM_{2.5}$ measurements from 2013 to 2016 to simulate historical data due to the lack of $PM_{2.5}$ measurements. This approach could improve the accuracy of historical $PM_{2.5}$ estimations by incorporating the interannual variations.

Our model exhibited suitable performance as well as the highest spatial resolution compared to those of other studies on $PM_{2.5}$ simulation in the Beijing-Tianjin-Hebei region (Lv et al., 2017; Wang et al., 2016). Our model performance was better than that of the multiple linear regression model that was applied by Wang et al. (Wang et al., 2016) (R^2 = 0.44–0.55), and the linear mixed-effects models applied by Zheng et al. (Zheng et al., 2016) (cv-R^2 = 0.77) and Lv et al. (2017) using Bayesian linear regression (R^2 = 0.75). Compared with models with similar spatiotemporal resolutions in Beijing, our model was also superior. The model performance was better than that of the mixed linear effect model (R^2 = 0.75–0.79) adopted by Xie et al. (2015) and was also better than that of the satellite-driven statistical model adopted by Liang et al. (2018) (cv-R^2 = 0.82). In contrast to the $PM_{2.5}$ simulation studies in China with low spatial resolution, our model exhibited better performance than that of the models by Zhan et al. (2017) and Ma et al. (2016) but worse performance than that of the model constructed by Li et al. (2017) using the neural network method. Although neural network model may exhibit excellent performance, the random forest model could avoid complex structures and consumed fewer computational resources. In addition, the differences between the modeling areas and basic modeling data may have led to different simulation performances. Our model also exhibited performance similar to that of a model applied to estimate the $PM_{2.5}$ concentrations in America (Hu et al., 2017; Di et al., 2016) at high spatiotemporal resolution (1 km × 1 km).

The model evaluation results for medium and high $PM_{2.5}$ concentrations were more accurate than those for low concentrations. The best performances of the model with regards to the seasons and regions (provincial and municipal level) occurred in the autumn and Beijing, respectively. However, the model performance was worse in the northern part of the study area. The latter may have been due to insufficient training samples with low $PM_{2.5}$ concentrations, which made our model more suitable for simulations of medium to high $PM_{2.5}$ concentrations. At the same time, the data fluctuations at low $PM_{2.5}$ concentrations during the simulation process were larger than those at high $PM_{2.5}$ concentrations, which could have caused the performance of the model to decline. Our research on regional performance achieved similar simulation results at higher spatial resolutions than those applied by Chen et al. (2018).

We considered the influence of meteorological hysteresis in the selection of the model variables. In the study by Chen et al. (2018), a nonlinear exposure-lag-response model was applied to estimate $PM_{2.5}$, but the model's verification results were poor in each province of China. To the best of our knowledge, we are the first to incorporate meteorological lag conditions into a random forest model. We used a meteorological factor with a two-day lag, which mainly considered the potential lag effects of the meteorological variables on the $PM_{2.5}$ estimation. In our research, the conditions of the meteorological lag factors greatly contributed to the model. The one-day lag of the atmospheric boundary layer was one of the major referenced lag variables. Adding a hysteresis effect variable increased the overall performance of the model by ~0.03. Similar to other studies (Pang et al., 2018; You et al., 2016; Lin et al., 2016; Lin et al., 2015), our study found that topographical factors and AOD data were two important contributors. In our research, land use data and traffic road data had small impacts on the model construction. Unlike the model described in Xiao et al. (2017), our model did not eliminate variables with low importance. An additional consideration was that the $PM_{2.5}$ concentration was also related to the variables with low importance. At the same time, retaining the variables with low importance did not cause a significant increase in the difficulty of model training due to the small amounts of model construction data. Key variables played a crucial role in the model. The hysteresis characteristics of meteorological effects should also be referenced in future research.

Our model used the highest spatial resolution of 0.01° × 0.01° in the study area. For parameter data with coarse resolution, we used bilinear interpolation to unify the resolution to 0.01° × 0.01°, which was commonly used during the exposure assessment process to meet the model precision requirements. The bilinear interpolation method produced a smoother interpolation, which has better performance than linear interpolation and nearest neighbor algorithms (Bovik, 2005). This method also has some biases; which have been reviewed in previous studies. These biases are inevitable in high-resolution simulation studies. Spatial resolutions of 4 km × 4 km (Lv et al., 2017) have been present in the studies on China's pollution level in the Beijing-Tianjin-Hebei region. In the small-scale area of Beijing, there has been a simulation at 1 km × 1 km spatial resolution (Liang et al., 2018). However, the spatial resolutions of $PM_{2.5}$ estimations were usually 10 km × 10 km or 0.1° × 0.1° at the country level (Ma et al., 2016; Zhan et al., 2017). More refined (high temporal resolution) and more accurate $PM_{2.5}$ population exposure measurements could provide fundamental data for the identification of population exposure effect-response relationships. Large amounts of data were incorporated into the model presented in this study, and the model was more complex and time consuming in large-scale areas. Part of our research included the simulation of historical $PM_{2.5}$ data, which utilized $PM_{2.5}$ measurements from 2013 to 2016. In previous studies (Ma et al., 2016; Liang et al., 2018), the use of short-term data to simulate long-term historical $PM_{2.5}$ data may have resulted in insufficient model training data, and it would not have been possible to accurately reflect changes in the data over a long temporal scale. In our study, $PM_{2.5}$ monitoring data from a longer period were used to simulate historical data, that is, to train a more generalized model. We preferred that the model adopted more data rules to improve the simulation results.

The $PM_{2.5}$ estimation showed significant differences in the seasonal and regional distributions. The $PM_{2.5}$ concentrations in the autumn and winter were higher in the southern part of the study area, which was similar to the results observed in other simulation studies. The concentration of $PM_{2.5}$ is low in summer due to the abundant precipitation, strong air convection, and high atmospheric boundary layer. In autumn and winter, pollutant emissions are due to the burning of coal in northern China, and pollution is aggravated by the control of stagnant weather conditions. In 2013 and 2014, the Beijing-Tianjin-Hebei region experienced high $PM_{2.5}$ concentrations, but these concentrations declined in 2015 and 2016. This phenomenon was closely related to the government's control over pollutant emissions. However, in the winter of 2016, $PM_{2.5}$ pollution was the most serious, mainly because only December was included for the winter period, and heavy pollution with long pollution times appeared from December 16 to 21. In most simulation studies, the simulated $PM_{2.5}$ concentration is usually lower than the monitored value, mainly because the monitoring sites are mostly distributed in urban areas, and model underestimation is a typical problem. Ma et al. (2014) predicted that the average annual $PM_{2.5}$ concentration in southern Beijing-Tianjin-Hebei in 2013 was approximately 90 μg/m^3, which is similar to the results presented in our research (96.4 μg/m^3). Our simulated $PM_{2.5}$ concentrations (94.03 μg/m^3) in 2014 were lower than those in Lv et al. (2016) (approximately 120 μg/m^3). Our results in winter (106.84 μg/m^3) were also lower than those in Lv et al. (180 μg/m^3). The reasons for the different $PM_{2.5}$ estimations in these studies may be due to the different data sources, data temporal ranges and model construction. The $PM_{2.5}$ simulations in our

study were close to the monitored concentrations. The $PM_{2.5}$ estimations in winter in the southern part of the study area were lower than the measurements, probably because the monitoring points tended to be clustered near the cities (higher concentrations), and our estimates were for the entire region. However, the $PM_{2.5}$ simulations in the northern part of the Beijing-Tianjin-Hebei region in the winter were lower than those in other seasons, which was similar to the results of the simulation study by Li et al. (2017) for 2014–2016. The low concentration of pollutants in the northern part of the study area was due to terrain control and low industrial emissions.

Our research has three advantages. First, we achieved high-resolution simulations of the $PM_{2.5}$ concentrations in the study area and achieved suitable performance. Second, we used a long-term $PM_{2.5}$ measurement training model to obtain a generalized model to estimate historical data. Third, we were the first to adopt the meteorological hysteresis effect into a random forest model. In our research, the meteorological lag conditions had important contributions to our model. However, our model still has two limitations. First, the model may have caused inaccurate $PM_{2.5}$ simulations in offshore areas due to the differences in the meteorological conditions between offshore and inland areas. This limitation could be improved by incorporating additional $PM_{2.5}$ monitoring values from offshore areas. Second, our model may have been more suitable for the simulation of medium and high $PM_{2.5}$ concentrations. Underestimation of the pollutant levels also occurred, which is a common issue in most studies.

5. Conclusions

The random forest model in this study was utilized to estimate the daily $PM_{2.5}$ concentrations at high spatial resolution ($0.01^{\circ} \times 0.01^{\circ}$) in the Beijing-Tianjin-Hebei region by incorporating AOD data, meteorological data, meteorological lag data, population data, elevation data, road data, land use coverage data, $PM_{2.5}$ concentrations from chemical models, and regional and seasonal dummy variables. The model showed suitable performance. The method in our study was a valid way to improve the spatiotemporal coverage of $PM_{2.5}$ concentrations. The study could supply $PM_{2.5}$ exposure assessments for long-term and short-term epidemiological studies: on the one hand, our model complemented the time series, providing exposure data for historical terms; on the other hand, the high spatial resolution modeling results provided exposure data for areas without monitoring sites in the Beijing-Tianjin-Hebei region. Exposure data with high spatial-temporal resolution reduce possible errors in exposure assessments for environmental epidemiological studies. In the future, we can consider more spatiotemporal factors related to $PM_{2.5}$ concentration over a larger research area to expand our study area and improve the modeling performance. Furthermore, validation of the modeling results in environmental epidemiology has rarely been tested. Studies assessing the applicability of modeling results with joint indicators are needed to provide a theoretical basis for the usage of these exposure data in environmental epidemiology.

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.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China [grant numbers 41701234]; the National Key Research and Development Program of China [grant numbers 2017YFC0211706].

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.105297.

References

Bovik, 2005. Handbook of Image and Video Processing, second ed. Elsevier Academic Press, pp. 35–36.
Cattani, G., Gaeta, A., Menno, Di, di Bucchianico, A., De Santis, A., Gaddi, R., Cusano, M., Ancona, C., Badaloni, C., Forastiere, F., Gariazzo, C., Sozzi, R., Inglessis, M., Silibello, C., Salvatori, E., Manes, F., Cesaroni, G., 2017. Development of land-use regression models for exposure assessment to ultrafine particles in Rome, Italy. Atmos. Environ. 156, 52–60.
Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60.
Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. Environ. Sci. Technol. 50 (9), 4712–4721.
Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Grazuleviciene, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Molter, A., Nador, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sμgiri, D., Tsai, M.Y., Yli-Tuomi, T., Varro, M.J., Vienneau, D., Klot, S., Wolf, K., Brunekreef, B., Hoek, G., 2012. Development of Land Use Regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46 (20), 11195–11205.
Fang, W., Yang, Y., Xu, Z., 2013. PM_{10} and $PM_{2.5}$ and health risk assessment for heavy metals in a typical factory for cathode ray tube television recycling. Environ. Sci. Technol. 47 (21), 12469–12476.
Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L., Strickland, M., Liu, Y., 2017. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51 (12), 6936–6944.
Hu, J., Chen, J., Ying, Q., Zhang, H., 2016a. One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system. Atmos. Chem. Phys. 16, 10333–10350.
Hu, L., Liu, J., He, Z., 2016b. Self-adaptive revised land use regression models for estimating $PM_{2.5}$ concentrations in Beijing, China. Sustainability 8 (8), 23.
Hu, X., Waller, L.A., Lyapustin, A., Wang, Y., Liu, Y., 2014. 10-year spatial and temporal trends of $PM_{2.5}$ concentrations in the southeastern US estimated using high-resolution satellite data. Atmos. Chem. Phys. 14 (12), 6301–6314.
Kheirbek, I., Wheeler, K., Walters, S., Kass, D., Matte, T., 2013. $PM_{2.5}$ and ozone health impacts and disparities in New York City: sensitivity to spatial and temporal resolution. Air Qual. Atmos. Health 6 (2), 473–486.
Kloog, I., Chudnovsky, A.A., Just, A.C., Nordio, F., Koutrakis, P., Coull, B.A., Lyapustin, A., Wang, Y.J., Schwartz, J., 2014. A new hybrid spatio-temporal model for estimating daily multi-year $PM_{2.5}$ concentrations across northeastern USA using high resolution aerosol optical depth data. Atmos. Environ. 95, 581–590.
Li, R., Cui, L., Li, J., Zhao, A., Fu, H., Wu, Y., Zhang, L., Kong, L., Chen, J., 2017a. Spatial and temporal variation of particulate matter and gaseous pollutants in China during 2014–2016. Atmos. Environ. 161, 235–246.
Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017b. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44 (23), 11985–11993.
Liang, F., Xiao, Q., Wang, Y., Lyapustin, A., Li, G., Gu, D., Pan, X., Liu, Y., 2018. MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China. Sci. Total Environ. 616–617, 1589–1598.
Lin, C., Li, Y., Yuan, Z., Lau, A.K., Li, C., Fung, J.C., 2015. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM2.5. Remote Sens. Environ. 156, 117–128.
Lin, C., Li, Y., Lau, A.K.H., Deng, X., Tse, T.K.T., Fung, J.C.H., Li, C., Li, Z., Lu, X., Zhang, X., Yu, Q., 2016. Estimation of long-term population exposure to PM2.5 for dense urban areas using 1-km MODIS data. Remote Sens. Environ. 179, 13–22.
Lu, X., Lin, C., Li, Y., Yao, T., Fung, J.C.H., Lau, A.K.H., 2017. Assessment of health burden caused by particulate matter in southern China using high-resolution satellite observation. Environ. Int. 98, 160–170.
Lv, B., Hu, Y., Chang, H.H., Russell, A.G., Bai, Y., 2016. Improving the accuracy of daily $PM_{2.5}$ distributions derived from the fusion of ground-level measurements with aerosol optical depth observations, a case study in North China. Environ. Sci. Technol. 50 (9), 4752–4759.
Lv, B., Hu, Y., Chang, H.H., Russell, A.G., Cai, J., Xu, B., Bai, Y., 2017. Daily, estimation of ground-level $PM_{2.5}$ concentrations at 4km resolution over Beijing-Tianjin-Hebei by fusing MODIS AOD and ground observations. Sci. Total Environ. 580, 235–244.
Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level PM2.5 in China using satellite remote sensing. Environ. Sci. Technol. 48 (13), 7436–7444.
Ma, Z., Hu, X., Sayer, A.M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., Liu, Y., 2016a. Satellite-based spatiotemporal trends in PM2.5 Concentrations: China, 2004-2013. Environ. Health Perspect. 124 (2), 184–192.
Ma, Z., Liu, Y., Zhao, Q., Liu, M., Zhou, Y., Bi, J., 2016b. Satellite-derived high resolution $PM_{2.5}$ concentrations in Yangtze River Delta Region of China using improved linear

mixed effects model. Atmos. Environ. 133, 156–164.
Pang, J., Liu, Z., Wang, X., Bresch, J., Ban, J., Chen, D., Kim, J., 2018. Assimilating AOD retrievals from GOCI and VIIRS to forecast surface PM2.5 episodes over Eastern China. Atmos. Environ. 179, 288–304.
Remer, L.A., Mattoo, S., Levy, R.C., Munchak, L.A., 2013. MODIS 3 km aerosol product: algorithm and global perspective. Atmos. Meas. Tech. 6 (7), 1829–1844.
Shi, Y., Lau, K.K., Ng, E., 2016. Developing street-level $PM_{2.5}$ and PM_{10} land use regression models in high-density Hong Kong with urban morphological factors. Environ. Sci. Technol. 50 (15), 8178–8187.
van Donkelaar, A., Martin, R.V., Spurr, R.J.D., Burnett, R.T., 2015. High-resolution satellite-derived $PM_{2.5}$ from optimal estimation and geographically weighted regression over North America. Environ. Sci. Technol. 49 (17), 10482–10491.
Wang, X., Guo, Y., Li, G., Zhang, Y., Westerdahl, D., Jin, X., Pan, X., Chen, L., 2016a. Spatiotemporal analysis for the effect of ambient particulate matter on cause-specific respiratory mortality in Beijing, China. Environ. Sci. Pollut. Res. 23 (11), 10946–10956.
Wang, Y., Jiang, H., Zhang, S., Xu, J., Lu, X., Jin, J., Wang, C., 2016b. Estimating and source analysis of surface PM2.5 concentration in the Beijing-Tianjin-Hebei region based on MODIS data and air trajectories. Int. J. Remote Sens. 37 (20), 4799–4817.
Xiao, Q., Wang, Y., Chang, H.H., Meng, X., Geng, G., Lyapustin, A., Liu, Y., 2017. Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sens. Environ. 199, 437–446.
Xie, Y., Wang, Y., Zhang, K., Dong, W., Lv, B., Bai, Y., 2015. Daily Estimation of ground-level $PM_{2.5}$ concentrations over Beijing using 3 km resolution MODIS AOD. Environ. Sci. Technol. 49 (20), 12280–12288.
Yang, J., Hu, M., 2018. Filling the missing data gaps of daily MODIS AOD using spatio-temporal interpolation. Sci. Total Environ. 633, 677–683.
Yanosky, J.D., Paciorek, C.J., Laden, F., Hart, J.E., Puett, R.C., Liao, D.P., Suh, H.H., 2014. Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. Environ. Health 13, 63.
Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y., 2018. Deep distributed fusion network for air quality prediction. In: Proceedings of the 24th SIGKDD Conference on Knowledge Discovery and Data Mining, London, United Kingdom, pp. 965–973.
You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., Wang, W., 2016. National-scale estimates of ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. Remote Sens. 8 (3), 184.
Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., Shen, X., Zhu, L., Zhang, M., 2017. Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139.
Zhang, T., Liu, G., Zhu, Z., Gong, W., Ji, Y., Huang, Y., 2016. Real-time estimation of satellite-derived $PM_{2.5}$ based on a semi-physical geographically weighted regression model. Int. J. Environ. Res. Public Health 13 (10), 13.
Zhao, C., Liu, Z., Wang, Q., Ban, J., Chen, N.X., Li, T., 2019. High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing-Tianjin-Hebei region. Atmos. Environ. 203, 70–78.
Zhao, P., Zhang, X., Xu, X., Zhao, X., 2011. Long-term visibility trends and characteristics in the region of Beijing, Tianjin, and Hebei, China. Atmos. Res. 101 (3), 711–718.
Zheng, Y., Zhang, Q., Liu, Y., Geng, G., He, K., 2016. Estimating ground-level PM2.5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. Atmos. Environ. 124, 232–242.