Science of the Total Environment 697 (2019) 134021

Contents lists available at ScienceDirect

Science of the Total Environment

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

Stacking machine learning model for estimating hourly $PM_{2.5}$ in China based on Himawari 8 aerosol optical depth data

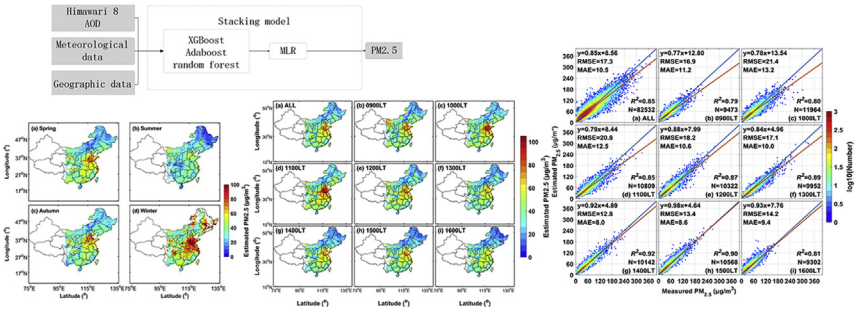
Jiangping Chen [a], Jianhua Yin [a,*], Lin Zang [b], Taixin Zhang [a], Mengdi Zhao [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan 430079, China

HIGHLIGHTS

- A stack model based on machine learning was developed to predict PM2.5 concentrations.
- Hourly PM2.5 concentrations in China were estimated in 2016 with R2 of 0.85.
- The effect of meteorological factors on PM2.5 was explored.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 April 2019
Received in revised form 3 July 2019
Accepted 19 August 2019
Available online 22 August 2019

Editor: Pingqing Fu

Keywords:
Air pollution
Himawari 8
Hourly $PM_{2.5}$
Stacking model

ABSTRACT

Aerosol optical depth (AOD) from polar orbit satellites and meteorological factors have been widely used to estimate concentrations of surface particulate matter with an aerodynamic diameter <2.5 μm ($PM_{2.5}$). However, estimations with high temporal resolution remain lacking because of the limitations of satellite observations. Here, we used AOD data with a temporal resolution of 1 h provided by a geostationary satellite called Himawari 8 to overcome this problem. We developed a stacking model, which contained three submodels of machine learning, namely, AdaBoost, XGBoost and random forest, stacked through a multiple linear regression model. Then, we estimated the hourly concentrations of $PM_{2.5}$ in Central and Eastern China. The accuracy evaluation showed that the proposed stacking model performed better than the single models when applied to the test set, with an average coefficient of determination (R^2) of 0.85 and a root-mean-square error (RMSE) of 17.3 μg/m^3. Model precision reached its peak at 14:00 (local time), with an R^2 (RMSE) of 0.92 (12.9 μg/m^3). In addition, the spatial and temporal distributions of $PM_{2.5}$ in Central and Eastern China were plotted in this study. The North China Plain was determined to be the most polluted area in China, with an annual mean $PM_{2.5}$ concentration of 58 μg/m^3 during daytime. Moreover, the pollution level of $PM_{2.5}$ was the highest in winter, with an average concentration of 73 μg/m^3.

1. Introduction

Particulate matter with an aerodynamic diameter <2.5 μm ($PM_{2.5}$) is associated with adverse health effects, particularly on the respiratory and cardiovascular systems (Chen et al., 2018; Lou et al., 2016; Schwartz et al., 2015). In China, rapid economic development and

* Corresponding author.
E-mail address: yinjianhua@whu.edu.cn (J. Yin).

https://doi.org/10.1016/j.scitotenv.2019.134021
0048-9697/

urbanisation have led to serious $PM_{2.5}$ pollution (Lu et al., 2017), which has attracted extensive attention. At the same time, particles and meteorological factors have a complex interaction, affecting the climate and the environment (Mao et al., 2018; Muhlbauer and Lohmann, 2008; Pan et al., 2018a, 2018b). In 2013, the National Air Quality Monitoring Network was established to monitor $PM_{2.5}$ in the entire country. However, the monitoring sites are mostly located in densely populated areas. To complement ground monitoring stations, satellite-retrieved aerosol optical depth (AOD) data have been widely used to estimate $PM_{2.5}$ concentration due to their advantage of wide coverage (Chu et al., 2016; Waller et al., 2014).

To date, an increasing number of statistical models have been established to predict $PM_{2.5}$ concentration, including the linear mixed-effects (LME) (Liu et al., 2017), generalised additive (Sun et al., 2015), geographically weighted regression (GWR) (Zhai et al., 2018), hierarchical and Bayesian (Lv et al., 2016) models. However, statistical models cannot effectively address the complex nonlinear relationship between dependent variables and predictors.

Compared with statistical models, a machine learning algorithm can better address the aforementioned nonlinear relationship, particularly for high-dimensional data, thereby providing it with better application prospect in air pollution prediction (Di et al., 2016; Hu et al., 2017). Deng et al. (2017) used a geographically weighted gradient boosting machine model to predict $PM_{2.5}$ concentration in China in 2014 and obtained a coefficient of determination (R^2) of 0.78. Li et al. (2018) and Hu et al. (2017) applied the random forest model to predict $PM_{2.5}$ exposure in China and the United States, respectively. To integrate the respective advantages of different models, combinatorial models have been gradually proposed and used to improve the accuracy of $PM_{2.5}$ estimation. Xiao et al. (2018) built a stacking model based on XGBoost, random forest and a generalised additive model to estimate daily $PM_{2.5}$ concentration in China and achieved an R^2 of 0.85. To date, however, most large-scale $PM_{2.5}$ prediction studies in China have been based on AOD data from the Moderate Resolution Imaging Spectroradiometer (MODIS) with a resolution of 1 day. This situation poses difficulty in meeting the demands of a detailed spatiotemporal monitoring of pollutants.

In the current study, a stacking model that combined AdaBoost, XGBoost and random forest was developed to estimate the hourly $PM_{2.5}$ concentrations from Himawari 8 AOD data in Central and Eastern China. Ground monitoring site data were firstly used to train each submodel, and then the multiple linear regression (MLR) model was applied to coalesce the results from the single models. Lastly, the spatial distribution of hourly $PM_{2.5}$ in China was mapped.

2. Materials and methods

2.1. Datasets

Himawari 8 AOD data, ground measurements of $PM_{2.5}$ concentrations, meteorological data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and geographic data were used in this study. All datasets spanned 1 year (from January 1, 2016 to December 31, 2016).

The Japan Meteorological Agency launched the weather satellite Himawari 8 on October 7, 2014; it carried the Advanced Himawari Imager with 16 bands (Kikuchi et al., 2018). Himawari 8 covers more than one-third of the Earth's surface, including the western Pacific, Oceania and Southeast Asia. In order to reduce the errors caused by haze and cloud on satellite AOD inversion (Mao et al., 2015), study chose Himawari 8 level 3 hourly AOD data confirmed high consistency by using ground-based measurements from the Aerosol Robotic Network (AERONET) (Wang et al., 2019; Zang et al., 2018).

Hourly $PM_{2.5}$ concentration data were obtained from the official website of the China Environmental Monitoring Centre (CEMC: http://106.37.208.233:20035). Meteorological data were obtained from ERA-Interim, one of the reanalysis datasets of ECMWF. The meteorological parameters adopted in this study included surface relative humidity (RH, %), boundary layer height (BLH, m), wind speed (u/v wind, m/s), surface pressure (SP, Pa) and temperature (TEMP, K). The normalised difference vegetation index (NDVI), which was used to approximate surface cover type in this study, was obtained from the MODIS 16-day NDVI production 'CMG 0.05 Deg 16 days NDVI' in 'MOD13C1/MYD13C1'. An NDVI higher than 0.4 typically represented a vegetation-covered area. In addition, we obtained the digital elevation model (DEM) data that covered China with a resolution of 90 m from the Consortium for Spatial Information (http://srtm.csi.cgiar.org/index.asp). The detailed information of the datasets used in this study is provided in Table 1.

2.2. Method

In previous $PM_{2.5}$ prediction studies, random forest, XGBoost and AdaBoost have exhibited good performance. Accordingly, the three machine learning models were stacked with MLR models in the current study to estimate and analyse hourly $PM_{2.5}$ concentrations in Central and Eastern China. The relevant mathematical theory and implementation process of the model are as follows.

2.2.1. Random forest

Originally proposed by Breiman (2001), the random forest algorithm is a combination of a series of decision trees in which each tree is an independent random sampling and all trees have the same distribution. The random forest algorithm gradually converges with an increase in the number of trees. This model is more accurate because of the injected randomness. The evaluation within the model is used to show the response to increasing the number of features used in splitting and to measure the importance of input variables (Breiman, 2001).

2.2.2. AdaBoost

AdaBoost is a boosting method that combines multiple weak classifiers into a strong classifier. From the beginning of training, the weights of the samples misclassified by the previous weak classifier are improved, and samples with updated weights are used to train the next weak classifier. With each new iteration, the model pays increasing attention to sample sets with high errors to improve accuracy (Drucker, 1997).

Table 1
Summary of the datasets used in this study.

Dataset	Variable	Unit	Temporal resolution	Spatial resolution	Source
$PM_{2.5}$	$PM_{2.5}$	$\mu g/m^3$	1 h	Site	CEMC
AOD	Satellite AOD	Unitless	1 h	0.18	Himawari 8
Meteorological factors	RH	%	6 h	0.125°	ECMWF
	SP	Pa			
	TEMP	K			
	U/V wind	m/s			
	BLH	m	3 h	0.125°	
Land	NDVI	Unitless	16 days	0.05°	MODIS
	DEM	m	Unavailable	90 m	NASA

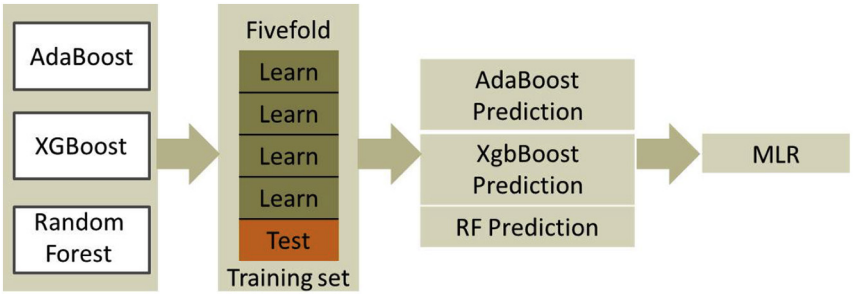


Fig. 1. Framework of stacking model training.

2.2.3. XGBoost

A gradient boosting decision tree (GBDT) is another boosting method. In contrast with AdaBoost, each classifier is designed to reduce the residual of the previous classifier. XGBoost, which is based on GBDT, allows the customization of loss functions. In the current machine learning competition, XGBoost has demonstrated excellent performance due to its high efficiency and impressive accuracy (Chen and Guestrin, 2016).

2.2.4. Stacking model and training

The three submodels described previously were stacked through the MLR model. The implementation is shown in Fig. 1. Firstly, this study adopted a fivefold crossover method to train each submodel with the training set. Then, the MLR model was trained with the predicted values of the test set from different models and observations. Lastly, the spatio-temporal continuous $PM_{2.5}$ could be estimated on the basis of the well-trained stacking model. In this study, R^2, mean absolute error (MAE) and root-mean-square error (RMSE) were used to evaluate the estimation performance of the stacking model.

3. Results and discussion

3.1. Effect of input variables on $PM_{2.5}$ estimation

The role of the input variables in each submodel was explored based on the analysis of feature importance. The statistical results are shown in Fig. 2. In this study, XGBoost, AdaBoost and random forest all conducted regression based on the tree-model, so obtained feature importance by computing the average change in the purity of node splitting in trees (known as Gini importance) (Randle, 2014). Despite the slight differences in the order of factor importance among different models, the top six factors were AOD, latitude, TEMP, BLH, longitude and RH. Longitude and latitude were fixed parameters, and the uncertainty of model estimation mostly originated from variable parameters (i.e. AOD, TEMP, BLH and RH).

AOD is the column integral of the atmospheric extinction coefficient, which is directly affected by the concentration of particulate matter and is the primary source of information for $PM_{2.5}$ concentration inversion. Zhang et al. (2018) indicated that satellite atmospheric aerosols are the main estimation factor of $PM_{2.5}$ concentration, and errors in satellite AOD data will lead to overestimation/underestimation of particulate matter concentration. BLH can be used to characterise the vertical diffusion capability of the atmosphere, and the low boundary layer strengthens atmospheric stability and increases air pollution. TEMP plays a decisive role in the development of the boundary layer (Huang et al., 2018; Liu et al., 2016). A low-temperature and high-humidity environment further promotes the formation of secondary aerosols. Previous studies have indicated that the generation of secondary aerosols is the primary reason for the explosive growth of particulate matter in China (Guo et al., 2014; Schwikowski et al., 2014). Meanwhile, temperature determines the amount of coal burned in North China during winter, which affects the anthropogenic emissions of $PM_{2.5}$. Therefore, we focused on the effects of AOD, TEMP, BLH and RH on $PM_{2.5}$ estimation in this study.

3.2. Model performance

The performance of the proposed and several commonly used $PM_{2.5}$ inversion models are summarized in Table 2. In model fitting, R^2 values ranged from 0.28 to 0.85, and MAE from 21.92 to 10.5 μg/m^3. Statistical results show that the MLR model performed the worst, with R^2 of 0.28 (Li et al., 2005). By introducing the spatial information, the GWR model significantly improved R^2 to 0.64 (Ma et al., 2014a). The LME model considering the random effect of AOD-$PM_{2.5}$ relationship changing with time had better performance than GWR, especially at a regional scale. Wang et al. (2017) estimated the $PM_{2.5}$ in Beijing-Tianjin-Hebei (BTH) region with a R^2 of 0.86. Artificial neutral network, as a new intelligent algorithm, is also applied to $PM_{2.5}$ concentration prediction.

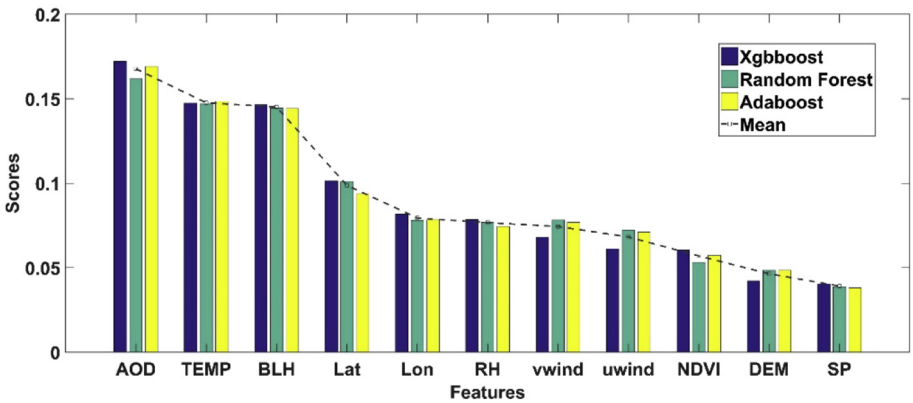


Fig. 2. Importance of features in different models.

Table 2
Summary of estimates of $PM_{2.5}$ concentration model on China.

Model	R^2	RMSE	MAE	Source of AOD	Reference
MLR	0.28	30.52	21.92	MODIS	(Li et al., 2005)
GWR	0.64	32.98	21.25	MODIS	(Ma et al., 2014b)
LME	0.78	27.99	18.67	MODIS	(Ma et al., 2016)
LME(BTH)	0.86	24.5	14.2	AHI	(Wang et al., 2017)
BPNN	0.47	25.96	18.06	MODIS	(Li et al., 2017)
PCN-GRNN	0.63	26.7	17.0	AHI	(Zang et al., 2018)
RF	0.82	19.6	12.2	AHI	This study
Adaboost	0.84	18.3	10.7	AHI	This study
XGBoost	0.84	18.1	11.4	AHI	This study
Stacking model	0.85	17.3	10.5	AHI	This study

However, back-propagation neural network (BPNN) was less accurate in predicting $PM_{2.5}$ concentration, with a R^2 of 0.47 (Li et al., 2017). Compared with BPNN, the combination of principal component analysis (PCA) and generalised regression neural network (GRNN) can more clearly show the relationship between $PM_{2.5}$ and multiple factors and had a 0.16 improvement on R^2 (Zang et al., 2018).

Among the machine learning models adopted in this study, XGBoost has the highest R^2 (0.84) and the lowest RMSE (18.1 μg/m^3), followed by with AdaBoost an R^2 of 0.84 and an RMSE of 18.3 μg/m^3, while the accuracy of RF is relatively low. As Wolpert (1992) indicated, the combination of multiple models can improve the robustness and generalization ability of models. The stacking model remarkably outperformed the individual models, with an R2 of 0.85. Comparisons with the previous inversion study of PM2.5 indicate that the stacking model proposed in this study is better than that of other models.

The comparison between the $PM_{2.5}$ concentration predicted by the stacking model and the measured values is provided in Fig. 3 (from 09:00 to 16:00, local time). At different times, the stacking model had a relatively high R^2 (i.e. 0.85 for all data and 0.79–0.92 for different hours), thereby indicating that the model achieved satisfactory temporal stability. However, the different R^2 values indicated that the performance of the stacking model was better at noon and in the afternoon during daytime. This finding is consistent with the results of Zang et al. (2019). By comparing the AOD data observed by Himawari 8 and AERONET, previous studies have found that the accuracy of Himawari 8 AOD data exhibits a significant fluctuation during the day and the highest accuracy at noon. In addition, the planetary boundary layer will extend upward due to the high temperature at noon, which will promote the mixing of particles in the vertical direction, thereby facilitating the estimation of fine particle concentration (Liu et al., 2016). High-accuracy AOD observation data and favourable atmospheric conditions ensure the effectiveness of model inversion.

Model performance in different seasons is shown in Fig. 4. Compared with that in the three other seasons, the model's performance in summer was poorer with an R^2 of 0.72. However, given the lower $PM_{2.5}$ levels in summer, the values of RMSE (11.0 μg/m^3) and MAE (7.5 μg/m^3) were relatively low. The model performed better in autumn and winter, with a high R^2 of 0.86 and RMSE (MAE) of 16.4 μg/m^3 (10.4 μg/m^3) and 21.4 μg/m^3 (12.6 μg/m^3), respectively. The model also achieved satisfactory performance in spring, with an R^2 value of 0.82. In summary, the model was stable during polluted seasons, particularly in autumn and winter.

To further explore the reason for the considerable seasonal differences in model performance, the correlations between the most important meteorological factors in the model (i.e. AOD, TEMP, BLH and RH)

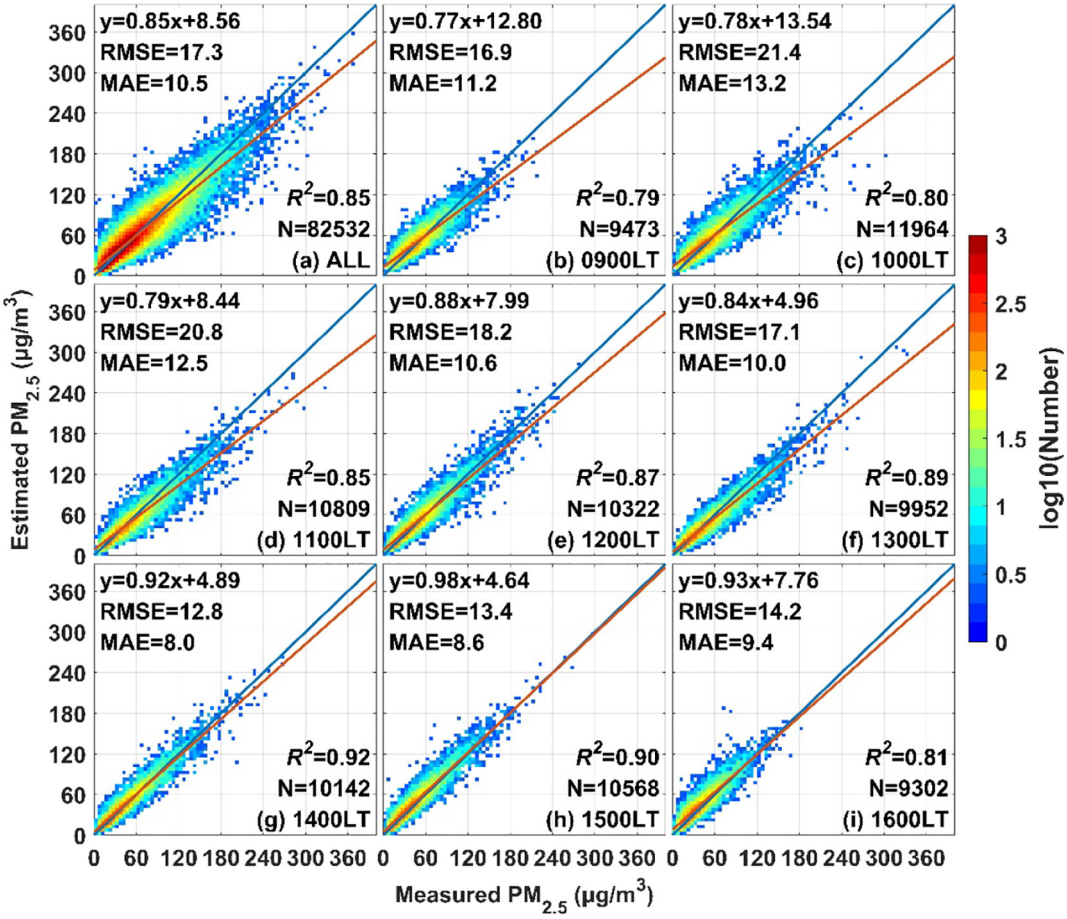


Fig. 3. Comparison between observed and predicted $PM_{2.5}$ values at different hours.

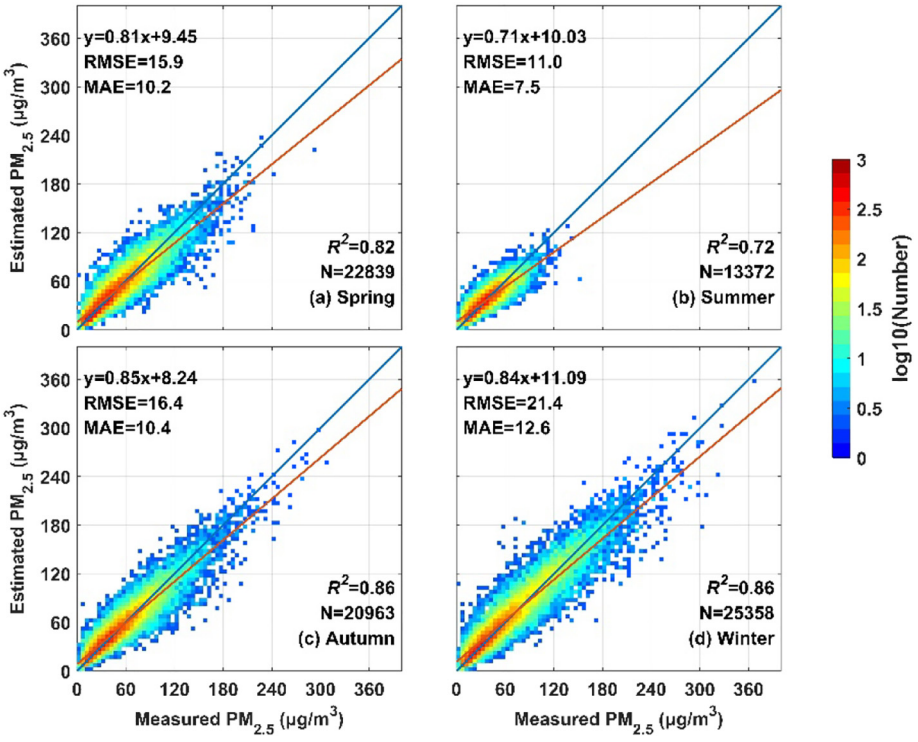


Fig. 4. Prediction of the model in different seasons.

and $PM_{2.5}$ in different seasons were investigated as shown in Fig. 5. Overall, the correlation between the factors and $PM_{2.5}$ was the highest in winter, followed by that in autumn, which guaranteed the enhanced performance of the proposed model in the two seasons. However, the significantly low correlation in summer would introduce considerable errors into ground-level $PM_{2.5}$ prediction. Moreover, cloud cover was greater in summer, thereby resulting in inaccurate and limited satellite observations, which increased the prediction errors of the model (Zang et al., 2019).

To evaluate the spatial performance of the proposed model, sites with less than five records were deleted, and their relative errors and R^2 were calculated as shown in Fig. 6(a) and (b). Model performance exhibited regional difference. In particular, the model performed best in the central and eastern regions of China (R^2 was typically >0.75), whereas the accuracy of the model in the southeastern coastal and western regions was relatively low, especially in the western regions (R^2 was mostly <0.65). However, >75% of the sites had an R^2 larger than 0.75, and about half of the sites had an R^2 higher than 0.85. In addition, >75% of the sites had a relative error of <35%. Overall, the site-based statistical results showed that the model was spatially stable.

Four key reasons could explain the spatial differences in model performance. Firstly, Central and Eastern China have more monitoring sites and data records, thereby resulting in a better training effect for the model. Less matchups will cause model overfitting and deteriorate estimation accuracy (Zhu et al., 2018). Secondly, different land cover types partly account for the difference. Western China is mostly covered with desert or ice–snow, and surface reflectivity is high, which increases uncertainty in AOD retrieval and introduces larger errors to model training (Guo et al., 2017). Thirdly, China's western region is located at the edge of the monitoring area of the Himawari 8 satellite, thereby resulting in low AOD coverage and precision. Lastly, as the column integral of particle extinction, AOD is influenced not just by fine particles, but by coarse

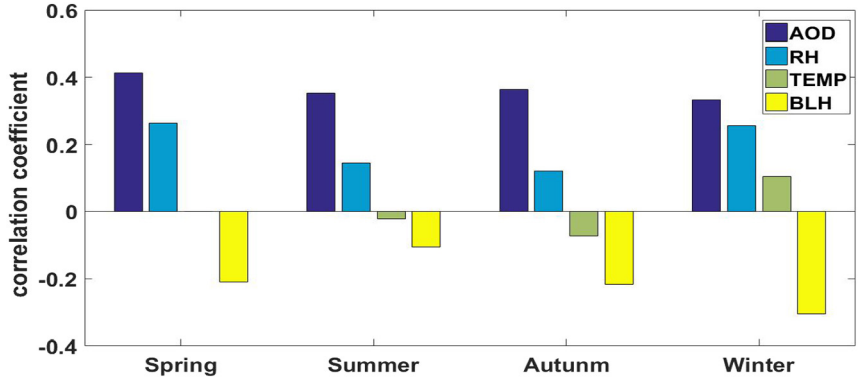


Fig. 5. Correlation between $PM_{2.5}$ and each parameter in different seasons.

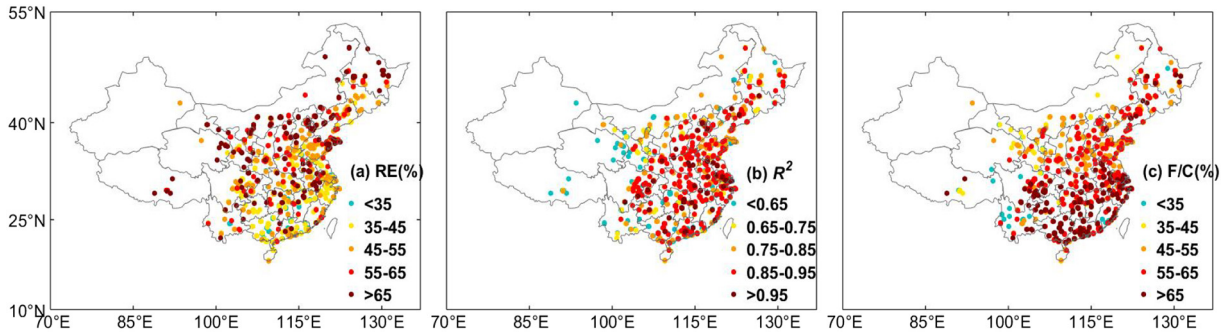


Fig. 6. Accuracy and the ratio of PM2.5/PM10 (denoted by F/C) distribution in sites.

particles, and the relationship between $PM_{2.5}$ and AOD changes with the ratio of $PM_{2.5}/PM_{10}$ (Agudelo-Castañeda et al., 2013; Yan et al., 2017). As shown in Fig. 6(c), the accuracy of the estimated PM2.5 based on sites is relatively consistent with the proportion of PM2.5, both increasing from the west to the east. Statistical results show that when $PM_{2.5}/PM_{10} < 50\%$, R^2 predicted by the model for $PM_{2.5}$ concentration at different sites was 0.77 on average, and R^2 value reached 0.81 when $PM_{2.5}/PM_{10} > 50\%$. This suggests the increase of the proportion of fine particles contributes to the improvement of prediction accuracy.

The central and eastern regions of China were the most polluted areas and the model performed the best in these areas. Therefore, we focused on the estimation of ground-level $PM_{2.5}$ in these areas.

3.3. Spatial and temporal distributions of $PM_{2.5}$

The near-surface $PM_{2.5}$ distribution with a spatial resolution of 0.05° in the central and eastern regions of China is shown in Fig. 7. In 2016, the annual average concentration of $PM_{2.5}$ was 40 μg/m^3, which were consistent the average in situ observation (39 μg/m^3).

Spatially, the North China Plain, which includes the Beijing–Tianjin–Hebei region and Shandong and Henan Provinces, was the most polluted region in China, with an estimated average annual $PM_{2.5}$ of 58 μg/m^3. On the one hand, rapid economic development and dense industrial distribution led to large emissions of pollutants. On the other hand, the Taihang Mountains in the western part of Hebei hindered the diffusion of fine particulate matter, thereby aggravating the pollution level in the region (Chen et al., 2018; Fu et al., 2014; Tao et al., 2012). The distribution of $PM_{2.5}$ was also high (average concentration of 56 μg/m^3) in the middle and lower reaches of the Yangtze River, where industrial activities, fuel burning, biomass burning and vehicle exhaust were the major contributors (Lou et al., 2016). In addition, the Sichuan Basin suffered from air circulation obstruction and atmospheric transmission due to its unique topography, thereby resulting in high $PM_{2.5}$ concentration (Xu et al., 2018). By contrast, areas with low $PM_{2.5}$ concentration were mostly located in sparsely populated areas, such as Inner Mongolia

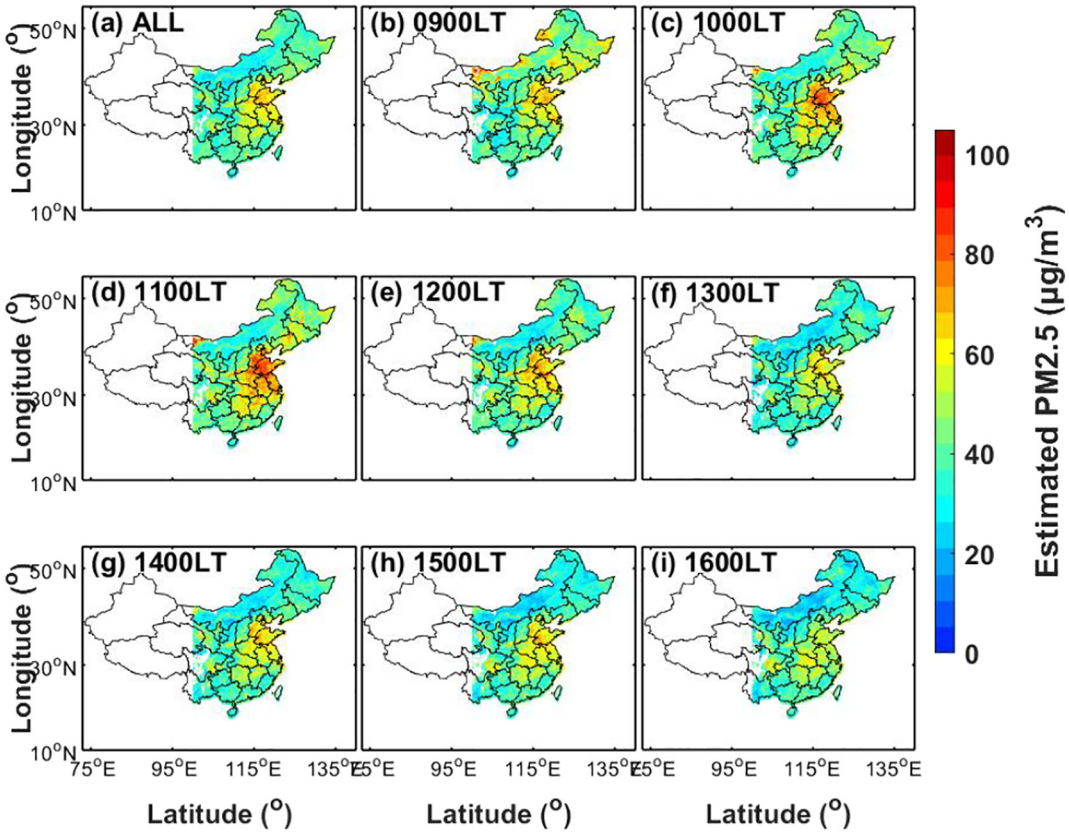


Fig. 7. Spatial and temporal distributions of $PM_{2.5}$ at different hours.

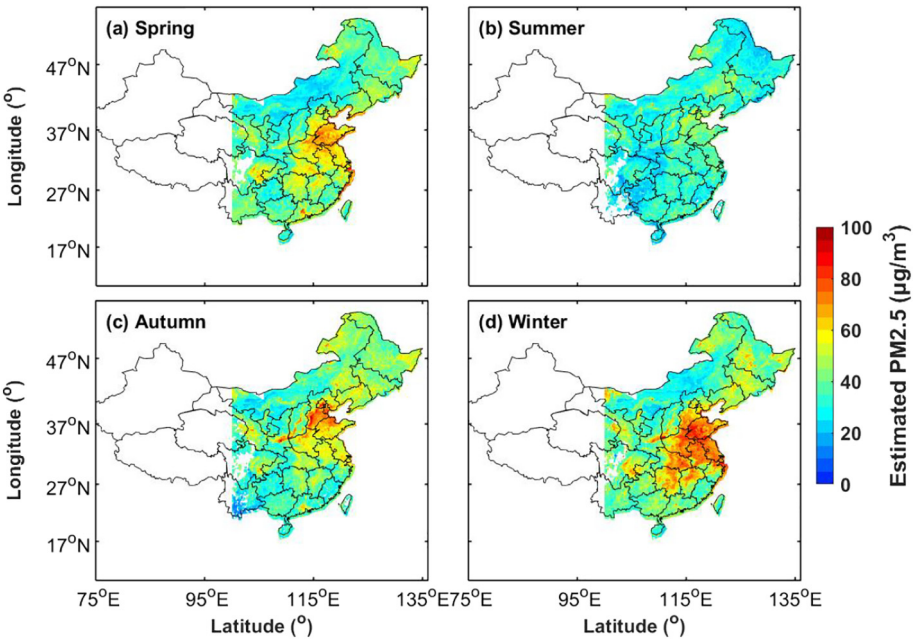


Fig. 8. Spatial and temporal distributions of $PM_{2.5}$ during different seasons.

and Southwest China, where $PM_{2.5}$ levels were typically below 35 μg/m^3.

In terms of time, $PM_{2.5}$ concentration also presented an apparent fluctuation, particularly in the eastern region of China. During daytime, the concentration of $PM_{2.5}$ at the national scale gradually increased from 09:00, peaked at 11:00 and then gradually decreased, which was consistent with ground measurements. In the most polluted North China Plain, $PM_{2.5}$ concentration gradually increased from 56 μg/m^3, reached a peak of 68 μg/m^3 and then declined to 47 μg/m^3. Meanwhile, $PM_{2.5}$ concentration ranged from 53 μg/m^3 to 64 μg/m^3 in the middle and lower reaches of the Yangtze River. By contrast, daily fluctuations were below 15 μg/m^3 in most areas with low pollution levels.

The spatial distribution of $PM_{2.5}$ concentrations in different seasons is shown in Fig. 8. Winter was the most polluted season, with an average $PM_{2.5}$ concentration of 54 μg/m^3. Summer was the cleanest season, with an average of 30 μg/m^3. These findings could be attributed to coal burning being the primary source of heat in North China, and winter being unfavourable for the diffusion of $PM_{2.5}$ (Zhang and Cao, 2015). The levels of particulate pollution were similar in spring and autumn, with average $PM_{2.5}$ concentrations of 36 μg/m^3 and 38 μg/m^3, respectively.

4. Conclusions

The spatial distribution of hourly $PM_{2.5}$ concentration is highly significant and necessary to understand the evolution of $PM_{2.5}$. Satellite-based AOD data are widely used in particulate matter estimation because of their wide coverage. However, the present research level in hours remains low. In this study, Himawari 8 AOD data were used to estimate hourly $PM_{2.5}$ concentration based on a proposed stacking model in Central and Eastern China. The results are as follows.

(1) The stacking model outperformed the individual models, achieving the highest R^2 of 0.85, followed by XGBoost (0.84), AdaBoost (0.84) and random forest (0.82). During daytime, the stacking model exhibited relatively high stability, with R^2 ranging from 0.79 to 0.92. The performance of the proposed model was better at noon and in the afternoon because of more accurate AOD observations and favourable weather conditions, which ensured the effectiveness of the proposed model.
(2) The model was more stable in polluted seasons, particularly in autumn and winter. The R^2 values in spring, summer, autumn and winter were 0.82, 0.72, 0.86 and 0.86, respectively. The results were determined by seasonal differences in the correlations between meteorological factors and $PM_{2.5}$ concentration. The statistical results showed that the correlation between the factors in autumn and winter was considerably higher than that in summer. In addition, the model demonstrated satisfactory spatial stability, and >75% of the sites had an R^2 higher than 0.75.
(3) The annual average concentration of $PM_{2.5}$ was 40 μg/m^3. The North China Plain and the middle and lower reaches of the Yangtze River delta were the most polluted areas, with average particulate matter concentrations of 58 μg/m^3 and 56 μg/m^3, respectively. The concentration of particulate matter presented evident daily and seasonal variations during daytime. The daily variation of $PM_{2.5}$ concentration was greater in areas with serious pollution, and the pollution level was most serious in winter.

In general, the estimation of near-surface $PM_{2.5}$ based on satellite AOD data is restricted by AOD quality. Meteorological factors also play an important role. In addition, different impacts of factors on $PM_{2.5}$ estimation under varying topography and climate conditions were not considered. In the future, we can study the spatial and temporal heterogeneity of factors for subregional prediction to further improve the accuracy and explanatory power of the model.

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

This study was supported by the National Key Research and Development Program of China (grant numbers 2018YFB10046 and 2017YFB0503604). The authors also grateful the CMA, Japan Aerospace Exploration Agency, ECMWF, Data Center of the US NASA, and USGS for supporting the data for this study.

References

Agudelo-Castañeda, D.M., Teixeira, E.C., Rolim, S.B.A., Pereira, F.N., Wiegand, F., 2013. Measurement of particle number and related pollutant concentrations in an urban area in South Brazil. Atmos. Environ. 70, 254–262. https://doi.org/10.1016/j.atmosenv.2013.01.029.

Breiman, L., 2001. Random forestsMachine Learning.

Chen, T., Guestrin, C., 2016. XGBoost. doi:https://doi.org/10.1145/2939672.2939785.

Chen, L., Gao, S., Zhang, H., Sun, Y., Ma, Z., Vedal, S., Mao, J., Bai, Z., 2018. Spatiotemporal modeling of PM2.5 concentrations at the national scale combining land use regression and Bayesian maximum entropy in China. Environ. Int. 116, 300–307. https://doi.org/10.1016/j.envint.2018.03.047.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., Xiang, H., 2016. A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. Atmosphere (Basel) https://doi.org/10.3390/atmos7100129.

Deng, X., Grieneisen, M.L., Shen, X., Zhu, L., Luo, Y., Zhang, M., Zhan, Y., Chen, H., 2017. Spatiotemporal prediction of continuous daily PM 2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139. https://doi.org/10.1016/j.atmosenv.2017.02.023.

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 Graphical abstract HHS Public Access. Environ. Sci. Technol. 50, 4712–4721. https://doi.org/10.1021/acs.est.5b06121.

Drucker, H., 1997. Improving regressors using boosting techniques. Proc. 14th Int. Conf. Mach. Learn.

Fu, G.Q., Xu, W.Y., Yang, R.F., Li, J.B., Zhao, C.S., 2014. The distribution and trends of fog and haze in the North China Plain over the past 30 years. Atmos. Chem. Phys. 14, 11949–11958. https://doi.org/10.5194/acp-14-11949-2014.

Guo, S., Shang, D., Zeng, L., Wu, Z., Molina, M.J., Hu, M., Du, Z., Zhang, R., Peng, J., Zamora, M.L., Shao, M., Zheng, J., 2014. Elucidating severe urban haze formation in China. Proc. Natl. Acad. Sci. 111, 17373–17378. https://doi.org/10.1073/pnas.1419604111.

Guo, J., Xia, F., Zhang, Y., Liu, H., Li, J., Lou, M., He, J., Yan, Y., Wang, F., Min, M., 2017. Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: implications for PM2.5 remote sensing. Environ. Pollut. 221, 94.

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, S., Guo, J., Lou, M., Yan, Y., Liu, S., Miao, Y., 2018. Unraveling the relationships between boundary layer height and PM2.5 pollution in China based on four-year radiosonde measurements. Environ. Pollut. 243, 1186–1195. https://doi.org/10.1016/j.envpol.2018.09.070.

Kikuchi, M., Murakami, H., Suzuki, K., Nagao, T.M., Higurashi, A., 2018. Improved hourly estimates of aerosol optical thickness using spatiotemporal variability derived from Himawari-8 geostationary satellite. IEEE Trans. Geosci. Remote Sens. 56, 3442–3455. https://doi.org/10.1109/TGRS.2018.2800060.

Li, C.C., Mao, J.T., Lau, A.K.H., Yuan, Z.B., Wang, M.H., Liu, X.Y., 2005. Application of MODIS satellite products to the air pollution research in Beijing. Sci. China Ser. D Earth Sci. 48, 209–219. https://doi.org/10.1360/05yd0395.

Li, T., Cao, W., Hamm, N.A.S., Li, S., Guo, Y., Abramson, M.J., Knibbs, L.D., Guo, J., Ren, H., Chen, G., 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. https://doi.org/10.1016/j.scitotenv.2018.04.251.

Liu, H., Lou, M., He, J., Li, Z., Zhang, Y., Bian, L., Zhai, P., Zhang, W., Yan, Y., Guo, J., Miao, Y., 2016. The climatology of planetary boundary layer height in China derived from radiosonde and reanalysis data. Atmos. Chem. Phys. 16, 13309–13319. https://doi.org/10.5194/acp-16-13309-2016.

Liu, M., Bi, J., Ma, Z., 2017. Visibility-Based PM2.5 Concentrations in China: 1957–1964 and 1973–2014. vol. 51 (acs.est.7b03468).

Lou, C.-R., Liu, H.-Y., Li, Y.-F., Li, Y.-L., 2016. Socioeconomic drivers of PM2. 5 in the accumulation phase of air pollution episodes in the Yangtze River Delta of China. Int. J. Environ. Res. Public Health 13, 928.

Lu, D., Xu, J., Yang, D., Zhao, J., 2017. Spatio-temporal variation and influence factors of PM2.5concentrations in China from 1998 to 2014. Atmos. Pollut. Res. 8, 1151–1159. https://doi.org/10.1016/j.apr.2017.05.005.

Lv, B., Hu, Y., Chang, H.H., Russell, A.G., Bai, Y., 2016. Improving the accuracy of daily PM2.5 distributions derived from the fusion of ground-level measurements with aerosol optical depth observations, a case study in North China. Environ. Sci. Technol. 50, 4752–4759. https://doi.org/10.1021/acs.est.5b05940.

Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y., 2014a. Supporting information-estimating ground-level PM2.5 in China using satellite remote sensing. Proc. Natl. Acad. Sci. 111, 12073–12078. https://doi.org/10.1073/pnas.1014723107.

Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y. %J E. Science, Technology, 2014b. Estimating Ground-level PM2. 5 in China Using Satellite Remote Sensing vol. 48, 7436–7444.

Ma, Z., Hu, X., Sayer, A.M., 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. https://doi.org/10.1289/ehp.1409481.

Mao, F., Duan, M., Min, Q., Gong, W., Pan, Z., Liu, G., 2015. Investigating the impact of haze on MODIS cloud detection. J. Geophys. Res. Atmos. 120, 12,212–237,247.

Mao, F., Pan, Z., Henderson, D.S., Wang, W., Gong, W., 2018. Vertically resolved physical and radiative response of ice clouds to aerosols during the Indian summer monsoon season. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2018.06.027.

Muhlbauer, A., Lohmann, U., 2008. Sensitivity studies of the role of aerosols in warm-phase orographic precipitation in different dynamical flow regimes. J. Atmos. Sci. 65, 2522–2542. https://doi.org/10.1175/2007jas2492.1.

Pan, Z., Mao, F., Wang, W., Logan, T., Hong, J., 2018a. Examining intrinsic aerosol-cloud interactions in South Asia through multiple satellite observations. J. Geophys. Res. Atmos. https://doi.org/10.1029/2017JD028232.

Pan, Z., Mao, F., Wang, W., Zhu, B., Lu, X., Gong, W., 2018b. Impacts of 3D aerosol, cloud, and water vapor variations on the recent brightening during the South Asian monsoon season. Remote Sens. https://doi.org/10.3390/rs10040651.

Randle, 2014. Python Machine Learning, Igarss. , p. 2014 https://doi.org/10.1007/s13398-014-0173-7.2.

Schwartz, J.D., Coull, B.A., Koutrakis, P., Melly, S.J., Zanobetti, A., Kloog, I., Shi, L., 2015. Low-concentration PM 2.5 and mortality: estimating acute and chronic effects in a population-based study. Environ. Health Perspect. 124, 46–52. https://doi.org/10.1289/ehp.1409111.

Schwikowski, M., Abbaszade, G., Ciarelli, G., Ho, K.-F., Baltensperger, U., Schnelle-Kreis, J., Zimmermann, R., Canonaco, F., Slowik, J.G., Haddad, I. El, Han, Y., Daellenbach, K.R., Piazzalunga, A., Cao, J.-J., Zotter, P., Prévôt, A.S.H., An, Z., Platt, S.M., Zhang, Y., Bozzetti, C., Szidat, S., Bruns, E.A., Wolf, R., Pieber, S.M., Huang, R.-J., Crippa, M., 2014. High secondary aerosol contribution to particulate pollution during haze events in China. Nature 514, 218–222. https://doi.org/10.1038/nature13774.

Sun, Q., Song, Y.-Z., Song, Y.-R., Li, Y., Peng, J.-H., Yang, H.-L., 2015. Estimating PM2.5 concentrations in Xi'an City using a generalized additive model with multi-source monitoring data. PLoS One 10, e0142149. https://doi.org/10.1371/journal.pone.0142149.

Li, T.W., Shen, H.F., Zeng, C., Yuan, Q.Q., Zhang, L.P., 2017. Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: methods and assessment. Atmos. Environ. 152, 477–489. https://doi.org/10.1016/j.atmosenv.2017.01.004.

Tao, M., Chen, L., Su, L., Tao, J., 2012. Satellite observation of regional haze pollution over the North China Plain. J. Geophys. Res. Atmos. 117. https://doi.org/10.1029/2012JD017915.

Waller, L.A., Lyapustin, A., Liu, Y., Wang, Y., Hu, X., 2014. 10-year spatial and temporal trends of PM2.5 concentrations in the southeastern US estimated using high-resolution satellite data. Atmos. Chem. Phys. 14, 6301–6314. https://doi.org/10.5194/acp-14-6301-2014.

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, 17. https://doi.org/10.3390/rs9080858.

Wang, W., Mao, F., Pan, Z., Gong, W., Yoshida, M., Zou, B., Ma, H., 2019. Evaluating aerosol optical depth from Himawari-8 with Sun photometer network. J. Geophys. Res. Atmos. 124, 5516–5538.

Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5, 241–259.

Xiao, Q., Chang, H.H., Geng, G., Liu, Y., 2018. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. Environ. Sci. Technol. 52, 13260–13269. https://doi.org/10.1021/acs.est.8b02917.

Xu, X., Li, Y., Zhang, L., Yin, X., Luo, L., Gui, K., Wang, H., Zhao, T., Gong, S., Guo, X., Zheng, Y., 2018. A modelling study of the terrain effects on haze pollution in the Sichuan Basin. Atmos. Environ. 196, 77–85. https://doi.org/10.1016/j.atmosenv.2018.10.007.

Yan, X., Shi, W., Li, Zhanqing, Li, Zhengqiang, Luo, N., Zhao, W., Wang, H., Yu, X., 2017. Satellite-based PM2.5 estimation using fine-mode aerosol optical thickness over China. Atmos. Environ. 170, 290–302. https://doi.org/10.1016/j.atmosenv.2017.09.023.

Zang, L., Mao, F., Guo, J., Gong, W., Wang, W., Pan, Z., 2018. Estimating hourly PM 1 concentrations from Himawari-8 aerosol optical depth in China. Environ. Pollut. 241, 654–663. https://doi.org/10.1016/j.envpol.2018.05.100.

Zang, L., Mao, F., Guo, J., Wang, W., Pan, Z., Shen, H., Zhu, B., Wang, Z., 2019. Estimation of spatiotemporal PM1.0 distributions in China by combining PM2.5 observations with satellite aerosol optical depth. Sci. Total Environ. 658, 1256–1264. https://doi.org/10.1016/j.scitotenv.2018.12.297.

Zhai, L., Li, S., Zou, B., Sang, H., Fang, X., Xu, S., 2018. An improved geographically weighted regression model for PM2.5 concentration estimation in large areas. Atmos. Environ. 181, 145–154. https://doi.org/10.1016/j.atmosenv.2018.03.017.

Zhang, Y.L., Cao, F., 2015. Fine particulate matter (PM 2.5) in China at a city level. Sci. Rep. 5, 14884. https://doi.org/10.1038/srep14884.

Zhang, K., Ma, F., Bilal, M., Guo, J., Lu, M., Zhang, Y., Qin, K., Zou, J., 2018. Estimating PM1 concentrations from MODIS over Yangtze River Delta of China during 2014–2017. Atmos. Environ. 195, 149–158. https://doi.org/10.1016/j.atmosenv.2018.09.054.

Zhu, Zerun, Xu, K., Shen, H., Gong, W., Zhang, T., Li, Z., Sun, K., Mao, F., Huang, Y., Wang, L., Zhu, Zhongmin, 2018. Estimation of ultrahigh resolution PM2.5 concentrations in urban areas using 160 m Gaofen-1 AOD retrievals. Remote Sens. Environ. 216, 91–104. https://doi.org/10.1016/j.rse.2018.06.030.