Supplementary Materials

A Robust Deep Learning Approach for Spatiotemporal Estimation of Satellite AOD and PM2.5

Lianfa Li [1,2]

[1] State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Datun Road, Beijing 100101, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

Number of pages: 25

Number of figures: 12

Number of tables: 3

Contents

1. Fusion of Daily Aqua and Terra AOD

The MODIS Aqua and Terra sensors usually provide different AOD images corresponding to their different overpass time. Due to different overpass time and meteorology (e.g. cloud coverage), these images usually have different missing patterns in space. It is a common practice to average Aqua and Terra AOD snapshots at two local time points (usually 10:30 am for Terra and 1:30 pm for Aqua) to get daily AOD averages. Sometimes only one of Aqua or Terra AOD is available and the other is missing [only one of Aqua or Terra AOD available at a spatial location/grid cell]. To better calculate daily averages and take full advantage of all AOD, Hu *et al*. (Hu et al., 2014) and Xiao *et al*. (Xiao et al., 2017) construct linear regression to associate Aqua AOD with Terra AOD. When one (Aqua or Terra AOD) is missing but the other is available, linear regression can be used to predict for the missing one.

For estimation of missing Aqua AOD (dependent variable), Terra AOD is the sole explanatory variable; for estimation of missing Terra AOD (dependent variable), Aqua AOD is the sole explanatory variable. Thus, using the data of both Aqua and Terra AOD available as the training samples, two regression models need to be trained to make prediction for the other missing AOD.

GAM is a non-linear regression model that had better performance than linear regression in estimation of Aqua or Terra AOD, as shown in sensitivity analysis. Thus, it was used in this paper to estimate the missing AOD. Then, the average over available or estimated AOD can be obtained to increase the size and spatial coverage of the training samples.

2. Covariates

2.1. Meteorological Measurement

Meteorological data came from daily ground observation values in China Mainland (Version 3.0) from the China Meteorological Data Service (http://data.cma.cn). The 2015 daily meteorological data [including daily air temperature (ºC), air pressure (hPa), relative humidity (%) and wind speed (m/s)] were collected from 824 national baseline meteorological stations in China, among which 27 were located in the Jing-Jin-Ji region.

2.2. Reanalysis Data

The reanalysis data provide reliable estimates of surface meteorological factors (e.g., air temperature, air pressure, relative humidity, wind speed and PBLH) or AOD at a regional scale (a coarse spatial resolution) (Parker, 2016). The reanalysis data of meteorological factors were gathered from the newest Goddard Earth Observing System-Forward Processing (GEOS-FP) dataset that was based on the Data Assimilation System (DAS) (ftp://rain.ucis.dal.ca/ctm/GEOS_0.25x0.3125_CH.d/GEOS_FP). Reanalysis data of MERRA2 AOD (as regional AOD) and PBLH were collected for 2015 (https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2). The dataset covers all of mainland China at a spatial resolution of 0.25º (latitude) × 0.3125º (longitude) and a temporal resolution of 3 h. Daily-level averages were obtained by averaging over 3 h reanalysis data.

2.3. Fusion of Meteorological Data

A GAM and autoencoder-based deep residual network were developed to fuse ground meteorological measurements and reanalysis data into high-resolution grid surfaces of the meteorological parameters, including air temperature, air pressure, relative humidity and wind speed. The results showed high CV R^2 (0.92 for air temperature and pressure; 0.94 for air pressure, 0.86 for relative humidity and 0.79 for wind speed). Ground measurement data were employed to obtain the fine local variability and reanalysis data at a regional scale to adjust the

final prediction. The prediction surfaces of meteorological parameters matched the high spatial (1km) and temporal (daily) resolution needed for MAIAC AOD imputation and PM$_{2.5}$ estimation and were used as the inputs to both models. For details about fusion of meteorological data, please refer to Li (2019) and Fang and Li (2019).

2.4. Coordinates and Elevation

The coordinates (latitude and longitude) and their derivatives (squares and products) were leveraged to capture complex spatial variation. Because the proposed models cannot directly embed spatial autocorrelation at the pixel level, they were used as a proxy for spatial autocorrelation.

The elevation data of 500-m spatial resolution were gathered from the Shuttle Radar Topology Mission (SRTM, https://www2.jpl.nasa.gov/srtm/), which was published in 2003.

2.5. Temporal Index

For the MAIAC AOD, the three continuous daily data of available MAIAC AOD were combined to train the models to predict the target (middle) day's AOD. Given a large sample size, the daily level models were trained using the combinational data from the samples of three continuous days. The temporal index (-1, 0, 1) was used to indicate different day indices for the samples.

For PM$_{2.5}$, due to its small sample size, a single base model was constructed using the Julian day as the temporal index to capture the temporal variation.

2.6. Match of the Measurement and Covariate Data

The measurement samples and the covariate data were matched by their corresponding spatial locations and measurement time (for spatial covariates such as elevation, just spatial locations were used for a match).

3. Autoencoder-Based Residual Network

This section provides technical details for the base model of the autoencoder-based residual network proposed in the main manuscript.

3.1. Autoencoder

As a special type of neural network, Autoencoder is designed to learn an efficient data compression or latent representation coding by the middle layer (Tschannen et al., 2018; Ya, 2019). In a typical auencoder, assuming a d-dimension input and output, $\mathbf{x}$, weight matrix, $\mathbf{W}$, bias vector, $\mathbf{b}$, the set θ of parameters, the layer index, L, and the activation function f, the mapping formula is given as follow:

$$\theta_{\mathbf{W},\mathbf{b}}(\mathbf{x}) : \mathbf{R}^d \to \mathbf{R}^d \tag{1}$$

$$\boldsymbol{\theta}_{\mathbf{W},\mathbf{b}}(\mathbf{x}) = f(\mathbf{W}^{(L)} f(\cdots f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})\cdots) + \mathbf{b}^{(L)}) \tag{2}$$

The optimal solution of the parameters, $\boldsymbol{\theta}_{w,b}$ can be solved by minimizing the loss function between $\mathbf{x}$ and its output estimates, $\mathbf{x}'$:

$$L = \frac{1}{2n}\left\|\mathbf{x} - \mathbf{x}'\right\|^2 = \frac{1}{2n}\left\|\mathbf{x} - \boldsymbol{\theta}_{\mathbf{W},\mathbf{b}}(\mathbf{x})\right\|^2 \tag{3}$$

3.2. Residual Connections and Backpropagation

Assume the shallow encoding layer, d, the latent coding layer, C, and its corresponding deep decoding layer, D. x_d and y_d denote the input and output of the encoding layer, d,

respectively, and x_D and y_D denote the input and output of the decoding layer, D, respectively. With the addition of residual identity mapping, the output of the deep decoding layer, D, is:

$$\mathbf{y}_D = \mathbf{x}_d + f(\mathbf{x}_D, \mathbf{W}_D) \tag{4}$$

$$\mathbf{x}_{D+1} = f(\mathbf{y}_D) \tag{5}$$

Given that D is the corresponding deeper layer for d, (4) can be rewritten as follows:

$$\mathbf{y}_D = \mathbf{x}_d + f(\ldots f(\mathbf{x}_d, \mathbf{W}_d) \ldots) \tag{6}$$

Based on automatic differentiation (Baydin et al., 2018) and Equation (6), the derivative of the loss function, L, for input of the shallow input, $\mathbf{x}_d$, can be:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{x}_d} &= \frac{\partial L}{\partial f_D(\mathbf{y}_D)} \frac{\partial f_D(\mathbf{y}_D)}{\partial \mathbf{y}_D} \frac{\partial \mathbf{y}_D}{\partial \mathbf{x}_d} \\
&= \frac{\partial L}{\partial f_D(\mathbf{y}_D)} \frac{\partial f_D(\mathbf{y}_D)}{\partial \mathbf{y}_D} (1 + \frac{\partial}{\partial \mathbf{x}_d} F(\ldots f_d(\mathbf{x}_d, \mathbf{W}_d) \cdots))
\end{aligned} \tag{7}$$

where $f_D(\mathbf{y}_D)$ and $f_d(\mathbf{y}_d)$ denote the activation functions for $\mathbf{y}_D$ and $\mathbf{y}_d$, respectively.

According to Equation (7), a residual connection makes one constant term and the regular terms available in the derivative of the comprehensive loss function in terms of the corresponding shallow layer's parameters. If an appropriate activation function such as ReLU is used to maintain $\partial f_D(\mathbf{y}_D)/\partial \mathbf{y}_D$ as linear or simple in Eq. (7), the error information of the deep layer [$\partial L / \partial f_D(\mathbf{y}_D)$] can be directly back-propagated to the shallow layer, $\mathbf{x}_d$, without affecting any weight layers, as has been demonstrated in the application of CNN (He et al., 2016a, b). In the proposed framework (Figure 1a), the residual connections are mapped in a nested way as shortcuts from multiple encoding layers to the corresponding decoding layers, which efficiently improves the back-propagations of the errors from the deep decoding layers to the shallow encoding layers in a parallel way. Thus, the introduction of residual connections can effectively solve the potential issues of saturation of gradients and degradation of accuracy in the regular neural network, as demonstrated in the fusion of the meteorological parameters (Fang and Li, 2019; Li, 2019).

3.3. ReLU Activation Function

Rectifier linear unit (ReLU) was used mostly in the hidden layers due to its efficient gradient propagation. ReLU is defined as the positive part of its argument:

$$f(x) = x^+ = \max(0, x) \tag{8}$$

where x is the input to a neuron. ReLU simplifies the calculation of $\partial f_D(\mathbf{y}_D)/\partial \mathbf{y}_D$ (=0 or 1) in Equation (7), thus maintaining the efficient back-propagation of the error during training.

4. Optimization of Hyperparamters

The gradient descent was used as the optimizer to find the (sub-) optimal solution for the models in training. A grid search was used to find an optimal solution for the hyperparameters, including the mini-batch size, number of decoding layers, number of nodes for each layer, learning rate, momentum, and activation functions in base models, as well as for sampling a proportion of the features and number of base models in bagging. A grid search is an exhaustive search for all possible values of parameters with high temporal complexity (Claesen and Moor, 2015). For two hyperparameters (i.e., a mini-batch size, a number of base models) that may have a local optimal value, the strategy of a binary search (Cormen et al., 2009) was used to decrease the temporal complexity from O(n) to O(log(n)) (n: number of options). To

ensure fairness of comparison, similar methods were used to find the optimal hyperparameters for XGBoost, a GAM and a regular neural network.

Specifically for this study, as an advanced method, the Adam optimization (Kingma and Ba, 2014) was used in model training. This method can adaptively update learning rate and the weights iteratively during learning and can achieve a better effect than classical stochastic gradient descent. The default configure of parameters (learning rate: 0.001, beta1: 0.9, beta2: 0.999) was used. The method of reducing learning rate when a metric has stopped improving was also used (patience: 20, minimum learning rate: 0.0001). The early stopping was also used to prevent overfitting (patience: 50). An optimal mini batch size is 512 for AOD and 1024 for PM$_{2.5}$. For this study, 100 base models sufficiently achieved the expected accuracy, as reported in the results.

5. XGBoost

As a scalable end-to-end tree boosting learning system, XGBoost (https://xgboost.readthedocs.io) is widely used to achieve state-of-the-art results in many domains (Chen and Guestrin, 2016). XGBoost is a sparsity-aware algorithm, and a cache-aware block structure is used for efficient tree learning. In this algorithm, the greedy heuristic algorithm or approximate algorithms can be used to construct the optimal trees according to the split score, i.e., the loss reduction after the split, based on the optimal weights and losses. For details, please refer to (Chen and Guestrin, 2016). In this paper, as a state-of-the-art machine learning method, XGBoost was tested for spatiotemporal estimation of PM$_{2.5}$ and compared with the proposed approach.

6. Processing of Extreme Values

For practical predictions, several trained base models may generate few extreme values beyond a normal range if the covariates have values beyond those of the training samples. Thus, for the MAIAC AOD, the valid range [0, 4] was used to constrain an individual base model's output; for PM$_{2.5}$, the outer fences (Iglewicz and Hoaglin, 1993) (defined as Q3+3×IQR, where Q3 is the third quartile and IQR is the interquartile range) were extracted from the daily training samples. For the valid daily PM$_{2.5}$ estimate, the lower bound was zero, and the upper bound was the maximum value among the outer fence and the maximum daily measured PM$_{2.5}$ concentration. If a predicted PM$_{2.5}$ was beyond the lower or upper bound, the bound would be used as an alternative to the predicted value.

References

Baydin, G.A., Pearlmutter, B., Radul, A.A., Siskind, J., 2018. Automatic differentiation in machine learning: a survey. Journal of Machine Learning Research 18, 1-43.

BMEPB, 2014. Main sources of PM2.5 in Beijing: vehicles, coal burning, industry, dust and neighboring cities. Beijing Municipal Ecological Environment Bureau.

BMEPB, 2018. A New Round of Beijing PM2.5 Source Analysis Officially Released. Beijing Municipal Ecological Environment Bureau.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, KDD.

Claesen, M., Moor, D.B., 2015. Hyperparameter Search in Machine Learning.

Cormen, H.T., Leiserson, E.C., Rivest, L.R., Stein, C., 2009. Introduction to algorithms (3rd ed.). MIT Press and McGraw-Hill.

Fang, Y., Li, L., 2019. Estimation of high-precision high-resolution meteorological factors based on machine learning. Journal of Geo-information Sciences (in Chinese) 21, 799-813.

He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J., 2016a. Deep Residual Learning for Image Recognition. Proc

Cvpr Ieee, 770-778.

He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J., 2016b. Identity Mappings in Deep Residual Networks. Lect Notes Comput Sc 9908, 630-645.

Hu, X.F., Waller, L.A., Lyapustin, A., Wang, Y.J., Al-Hamdan, M.Z., Crosson, W.L., Estes, M.G., Estes, S.M., Quattrochi, D.A., Puttaswamy, S.J., Liu, Y., 2014. Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment 140, 220-232.

Iglewicz, B., Hoaglin, C.D., 1993. How to detect and handle outliers, in: Mykytka, F.E. (Ed.), The ASQ Basic References in Quality Control: Statistical Techniques. American Society for Quality, Milwaukee.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, L., 2019. Geographically weighted machine learning and downscaling for high-resolution spatiotemporal estimations of wind speed. Remote Sens 11, 1378.

Parker, W.S., 2016. REANALYSES AND OBSERVATIONS What's the Difference? B Am Meteorol Soc 97, 1565-+.

Tschannen, M., Bachem, O., Lucic, M., 2018. Recent Advances in Autoencoder-Based Representation Learning, Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montreal, Canada.

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 Sensing of Environment 199, 437-446.

Ya, J.Y., 2019. Autoencoder node saliency: Selecting relevant latent representations. Pattern Recognition 88, 643-635.

Table S1. Descriptive Statistics for the covariates and target variables (2015).

Group	Variable	Correlation with MAIAC AOD	Correction with PM$_{2.5}$	Annual Mean	Annual Sd.[a]	Summer Mean	Summer Sd.	Winter Mean	Winter Sd.
Target variable	MAIAC AOD	1	0.39	0.38	0.17	0.55	0.23	0.43	0.19
	PM$_{2.5}$ (µg/m^3)	0.39	1	80.2	70.46	57.16	34.26	121.91	97.59
Converted (ground) AOD	MAIAC AOD by PBLH and RH[b]	0.59	0.54	0.0014	0.0016	0.0009	0.0007	0.002	0.002
Meteorology	Air temperature ((°C))	0.29	-0.33	12.56	9.06	23.18	1.94	0.37	2.02
	Wind speed (m/s)	-0.23	-0.21	19.28	7.59	17.56	5	18.43	8.24
	Air pressure (hpa)	0.42	0.28	1001.97	25.44	994.93	24.92	1008.5	24.81
	Relative humidity (%)	0.39	0.19	59	17	66	12	54	16
Reanalysis data	PBLH	0.09	-0.25	844.95	495.13	1242.2	292.92	359.85	210.06
	MERRA2 AOD	0.11	-0.08	0.12	0.13	0.18	0.21	0.08	0.04
Coordinate	Latitude (°)	-0.44	-0.16	39.32	1.13	39.31	1.13	39.31	1.13
	Longitude (°)	-0.43	-0.07	116.23	1.1	116.23	1.11	116.23	1.11
Elevation	Elevation (meter)	-0.43	-0.21	123.86	213.22	122.21	212.83	122.21	212.83

Note: [a]. Sd. Standard deviation. [b]. MAIAC AOD by PBLH and RH: MAIAC AOD corrected by PBLH and relative humidity.

Table S2. The performance of the re-trained models for MAIAC AOD .

Date	Training R²	Training RMSE	Test R²	Test RMSE
04/20/2015	0.97	0.071	0.97	0.071
10/20/2015	0.97	0.13	0.96	0.14

Table S3. Pearson's correlation between uncertainty (standard deviation) of grid surfaces of PM2.5 and the covariates.

Covariate	Mean correlation for a year
MAIAC AOD	0.46
Corrected MAIAC AOD	0.44
Air temperature	0.27
Wind speed	-0.10
Air pressure	0.20
Relative humidity	0.17
PBLH	0.03
MERRA2 AOD	-0.005
Latitude	-0.07
Longitude	-0.38
Elevation	-0.20

(**a**) Percentage of PM2.5 sources in 2014 (BMEPB, 2014)

(**b**) Percentage of PM2.5 sources in 2018(BMEPB, 2018)
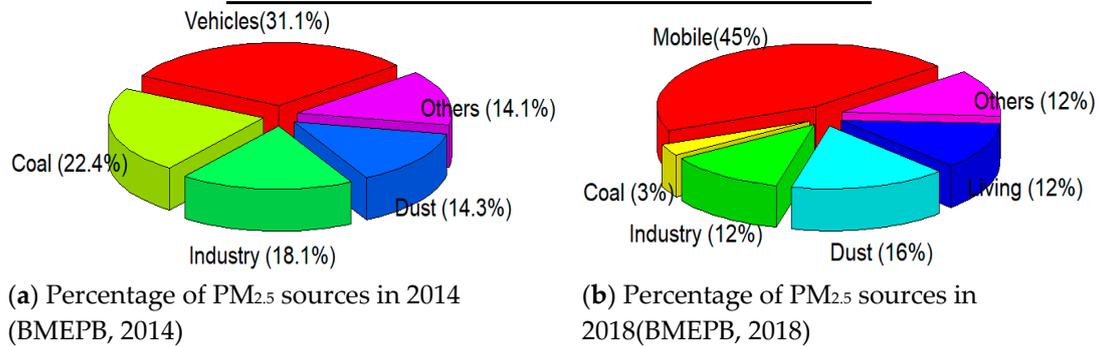
Figure S1. Distribution of 2014 vs. 2015 PM2.5 emission sources in Beijing.
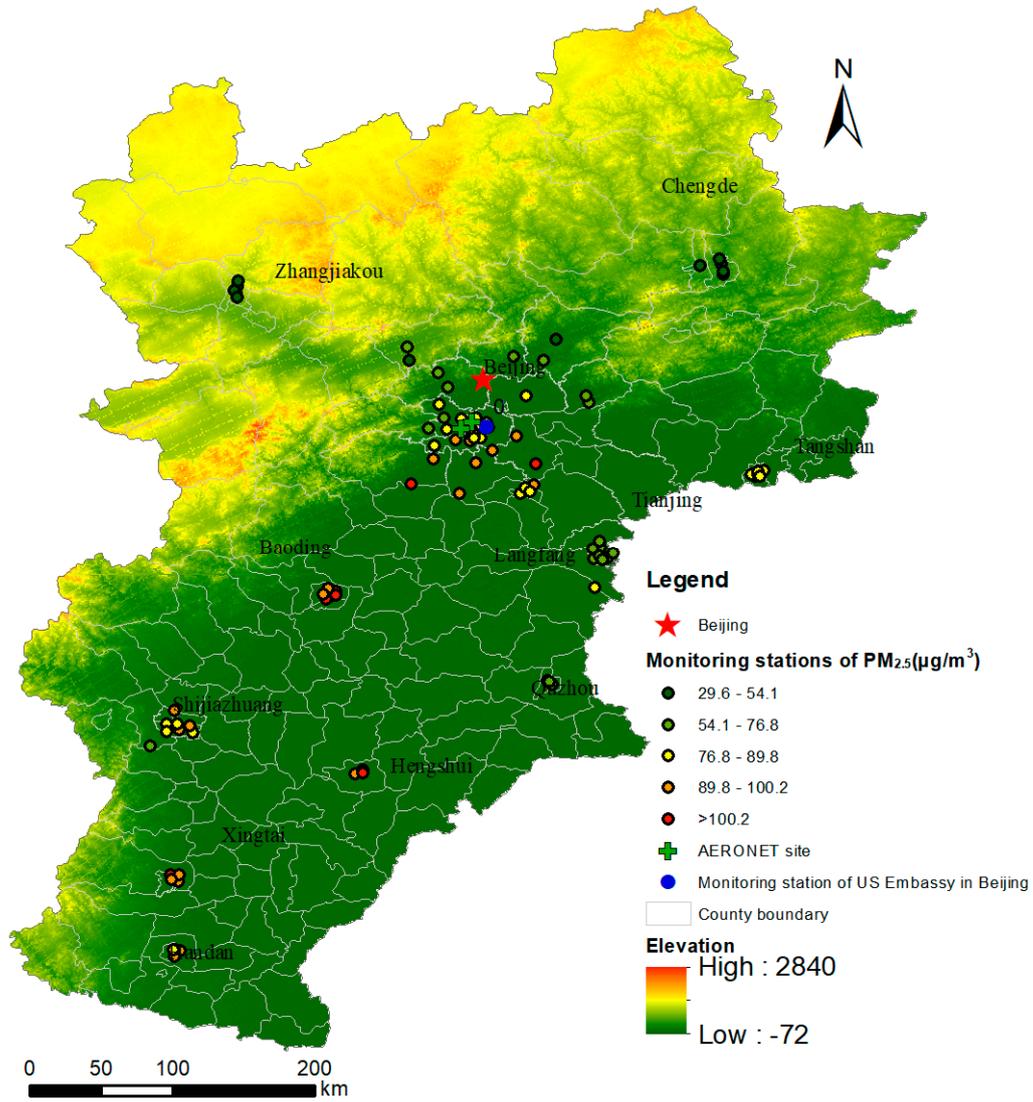
Figure S2. Study region (covering most of the Jing-Jin-Ji metropolitan area) with distributions and 2015 annual averages of PM$_{2.5}$ ground monitoring stations, AOD sites of AERONET, and monitoring stations of the US embassy in Beijing for independent test.
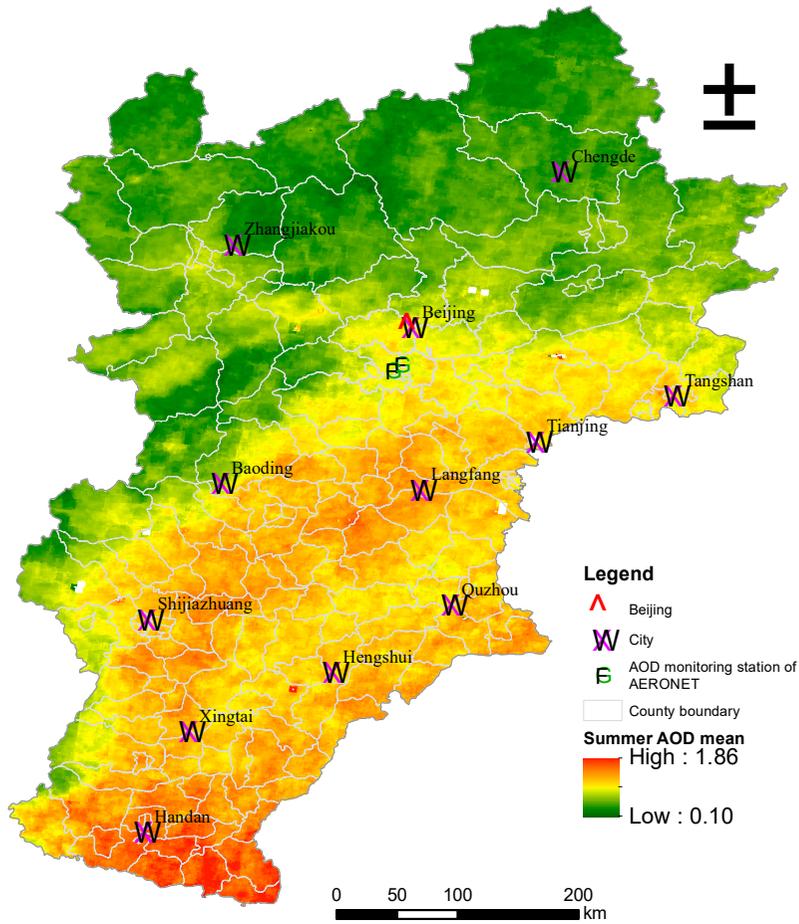
(**a**) AOD mean for summer of 2015
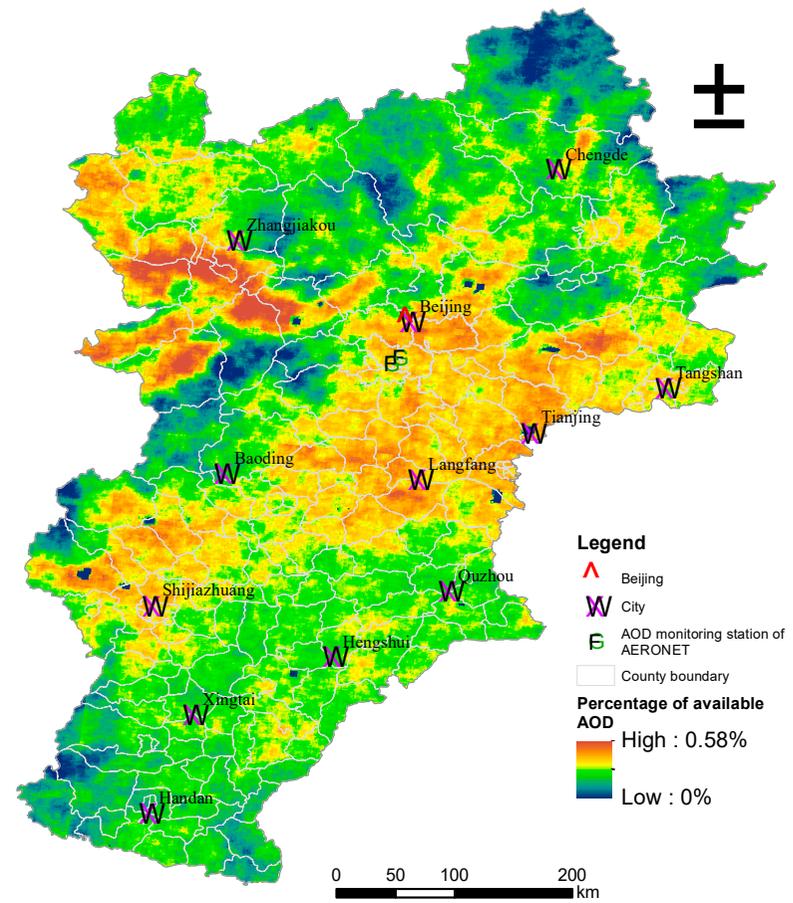
(**b**) Percentage of available MAIAC AOD for summer of 2015
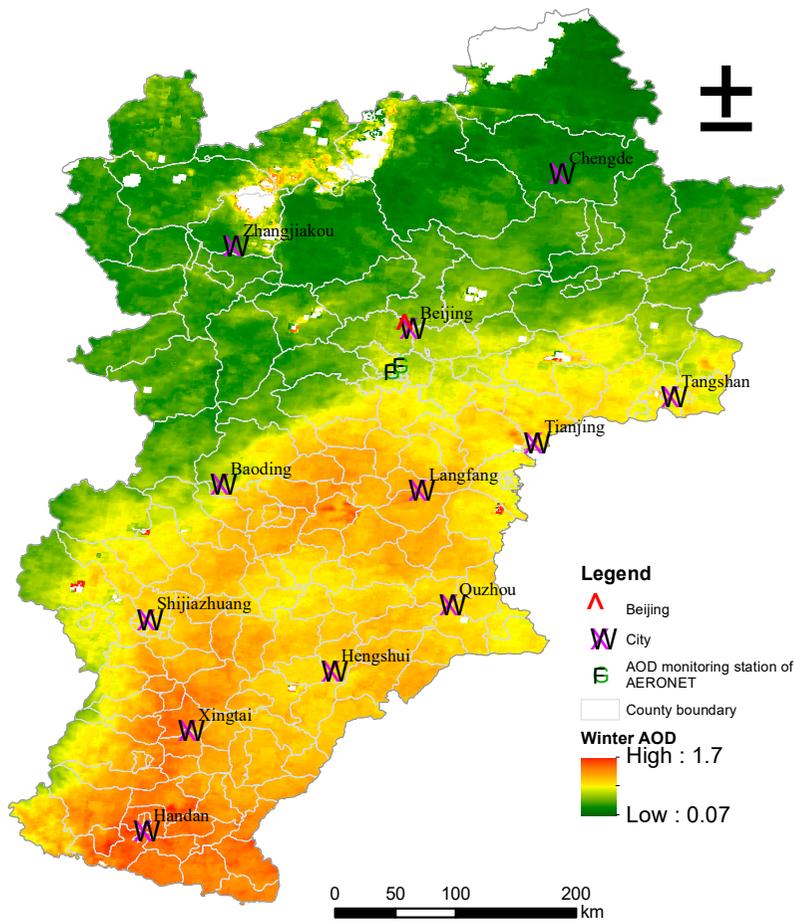
(**c**) AOD mean for winter of 2015
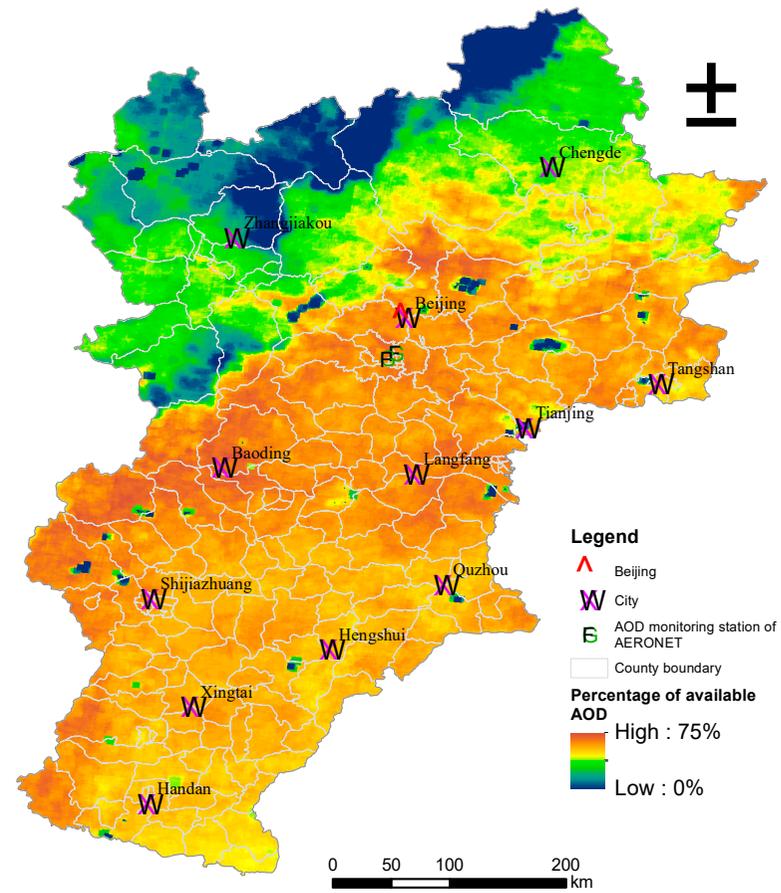
(**d**) Percentage of available MAIAC AOD for winter of 2015

Figure S3. Statistics for summer (a and b) and winter (c and d) of 2015.

(**a**) Loss curve for 04/20/2015 (spring)

(**b**) Loss curve for 07/20/2015 (summer)

(**c**) Loss curve for 10/20/2015 (autumn)

(**d**) Loss curve for 12/01/2015 (winter)
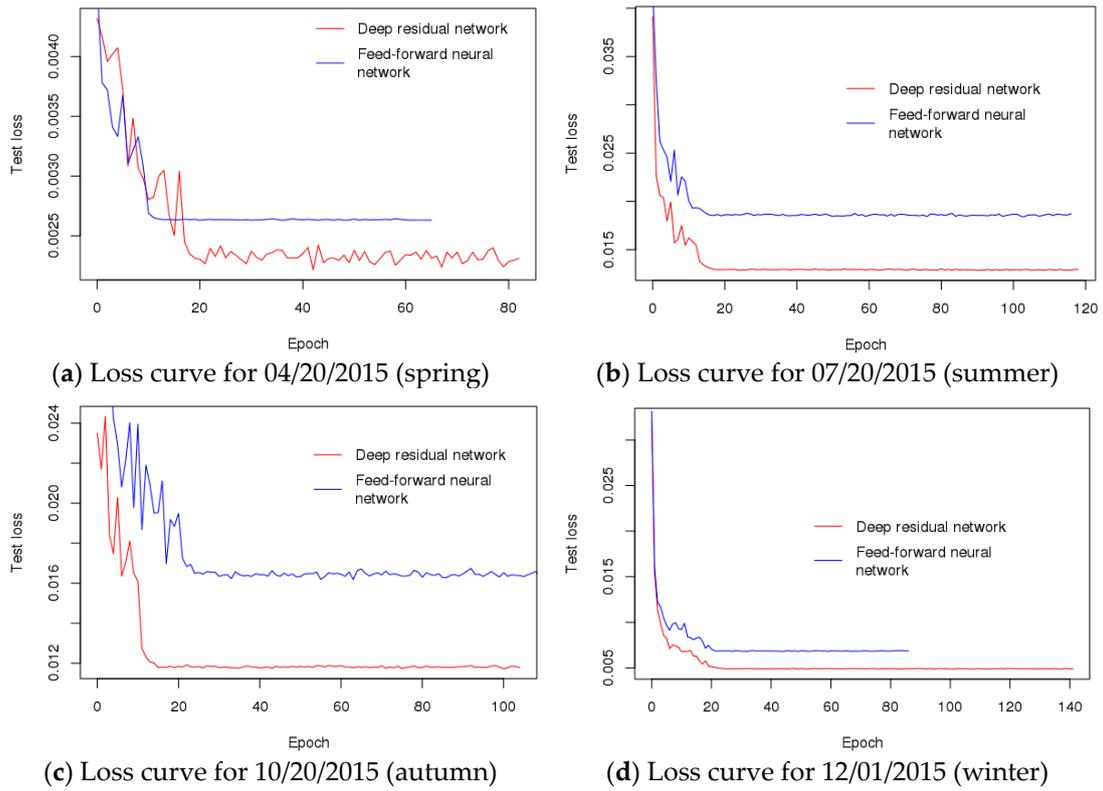
Figure S4. Learning curves of test loss of four typical seasonal days of 2015 for MAIAC AOD imputation.

(**a**) Spring day (04/20/2015)

(**b**) Summer day (07/20/2015)

(**c**) Autumn day (10/20/2015)

(**d**) Winter day (12/01/2015)
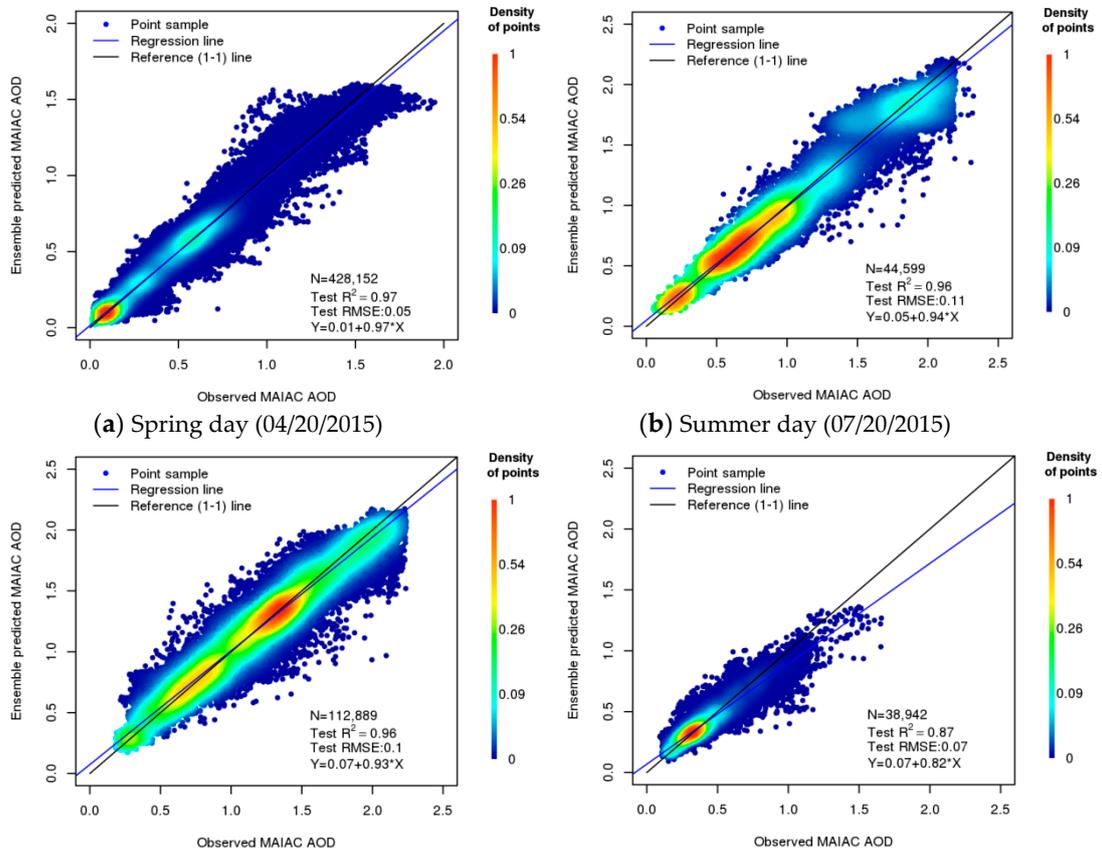
(**c**) Autumn day (04/20/2015)   (**d**) Winter day (07/20/2015)

Figure S5. Plots of observed vs. imputed daily MAIAC AOD in the test for four typical seasonal days in 2015.

(**a**) Spring day (04/20/2015)   (**b**) Summer day (07/20/2015)

(**c**) Autumn day (10/20/2015)   (**d**) Winter day (12/01/2015)
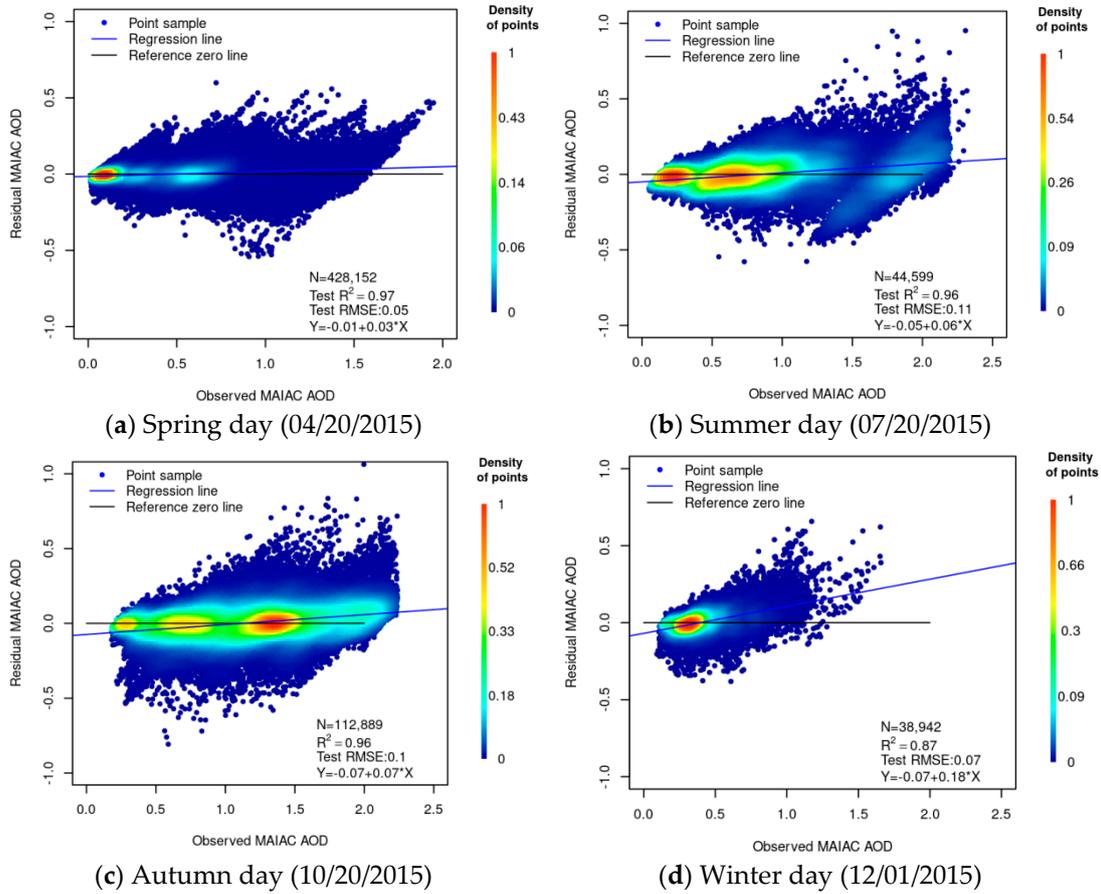
Figure S6. Plots of observed vs. residual MAIAC AOD for four typical seasonal days in 2015.

(**a**)   (**b**)
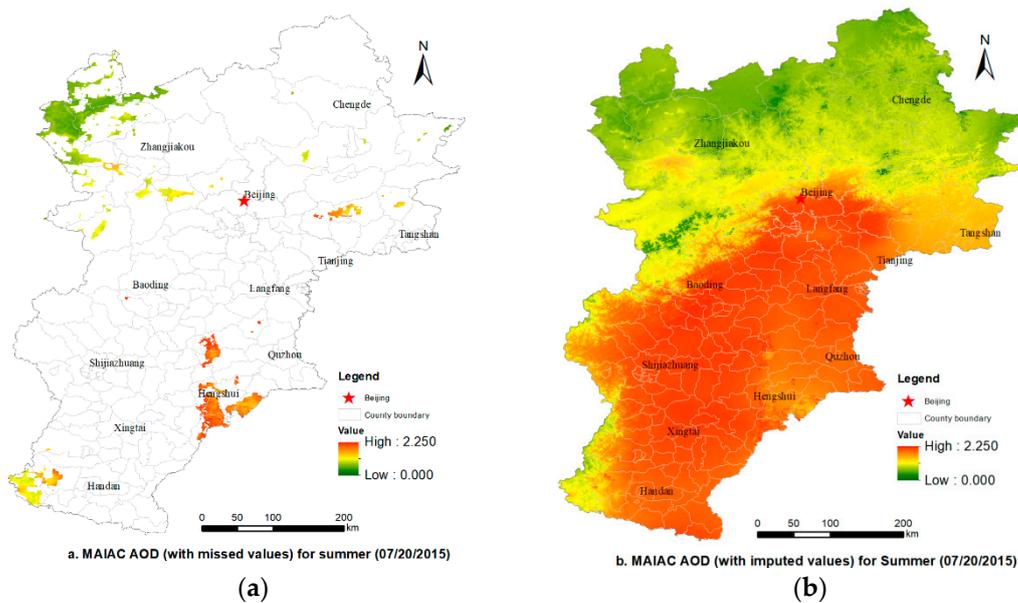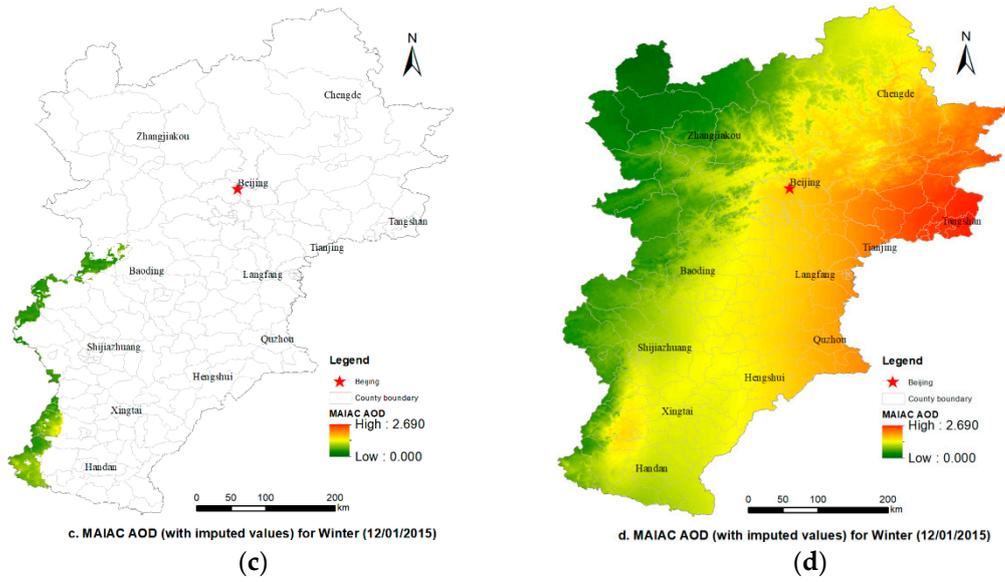
Figure S7. Original (a and c) and imputed MAIAC AOD (b and d) of the study region for two seasonal days (a and b: summer; c and d: winter) in 2015.
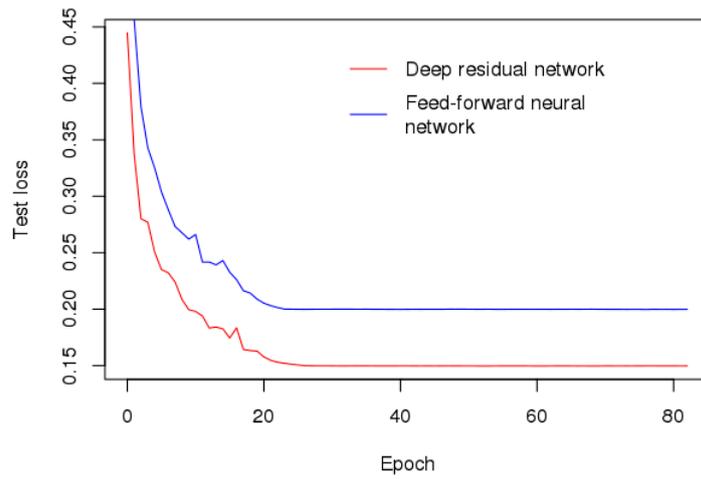
Figure S8. Learning curves of validation loss for PM$_{2.5}$ estimation.
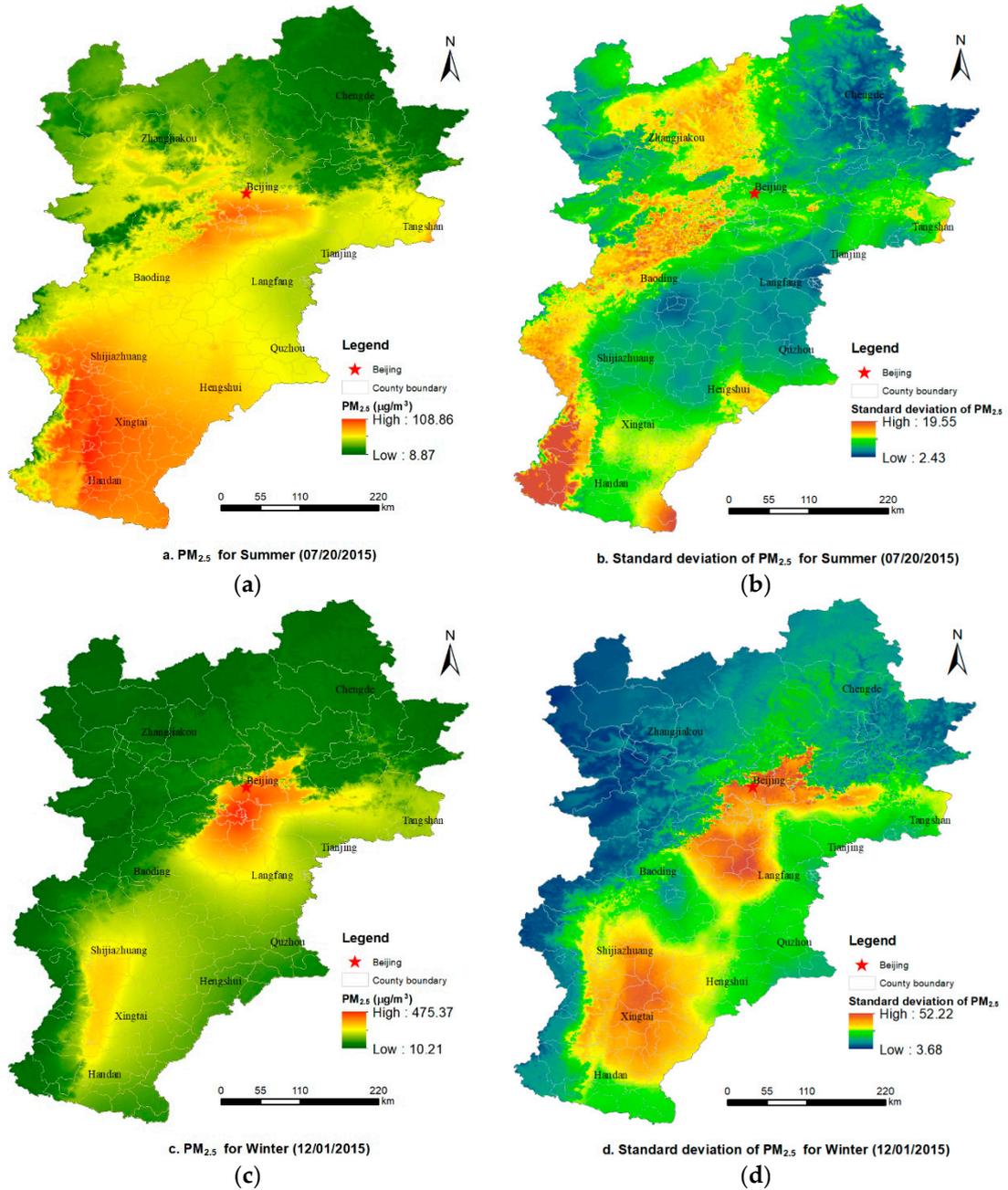
Figure S9. Grid surfaces of predicted PM$_{2.5}$ (μg/m^3) (a and c) and its standard deviation (b and d) for two seasonal days (a and b: summer; c and d: winter) in 2015.
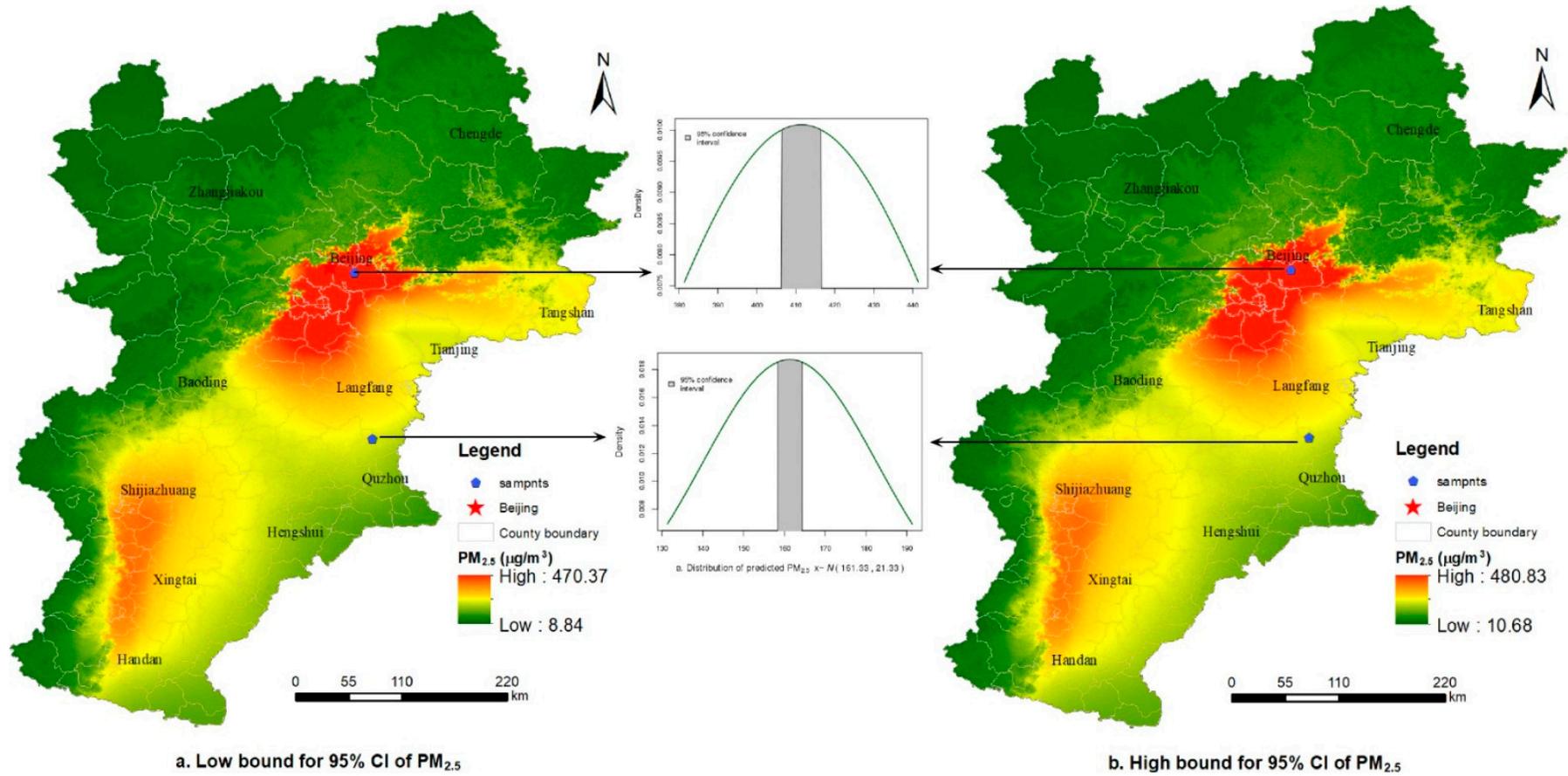
Figure S10. Images for low (**a**) and high (**b**) bounds of the 95% confidence interval of predicted PM$_{2.5}$ (winter day of /12/01/2015).
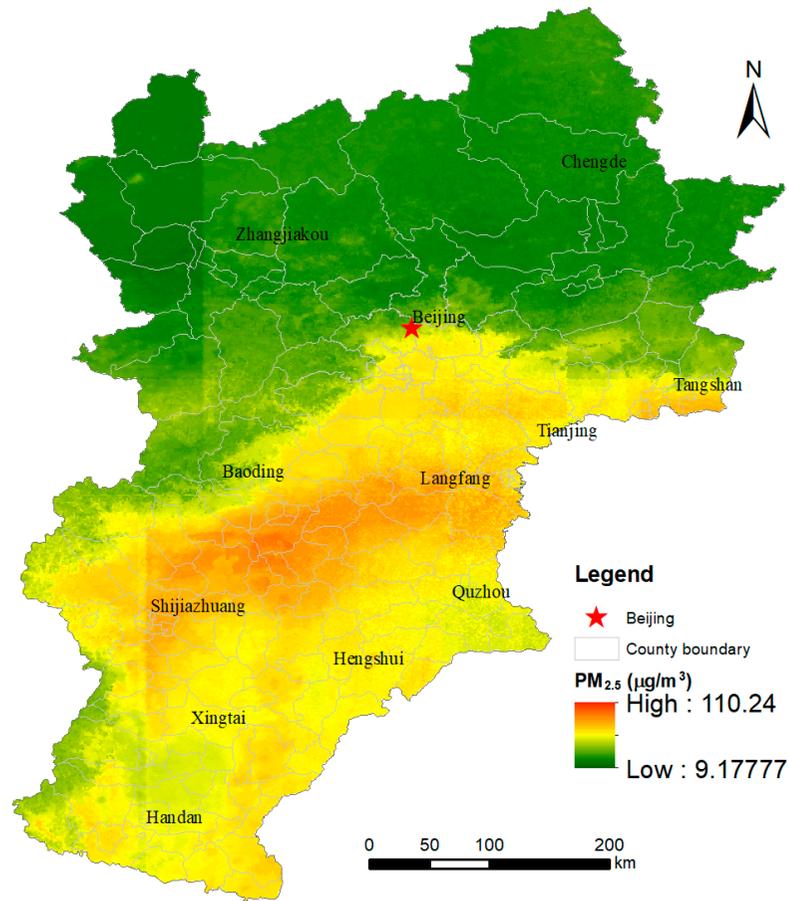
a. PM$_{2.5}$ for Spring (04/20/2015)

(**a**)
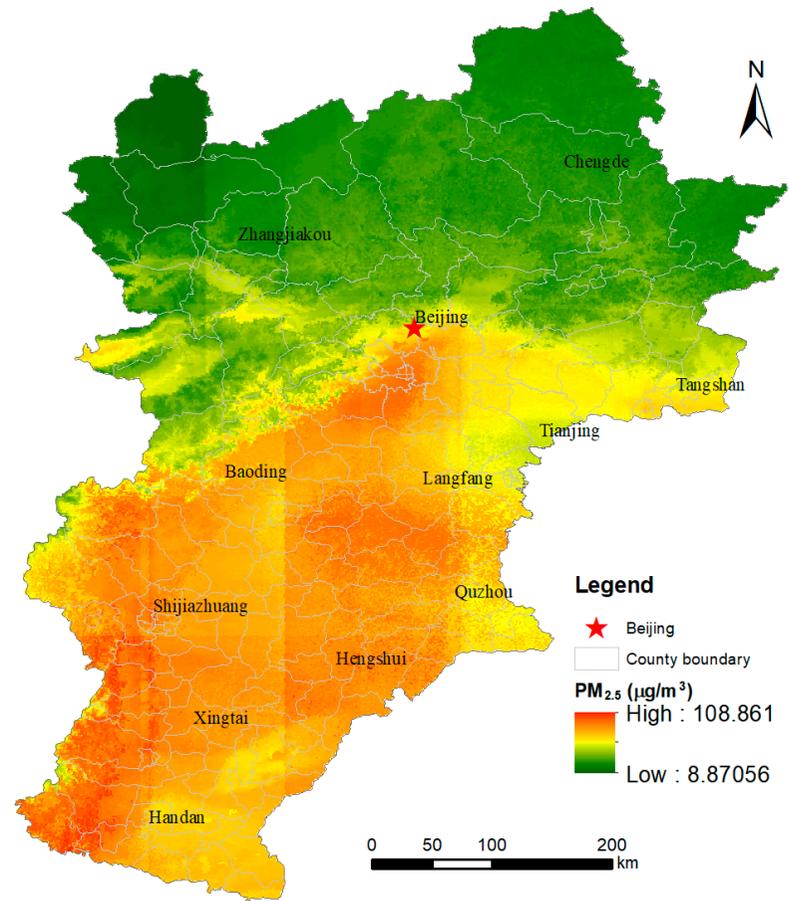
b. PM$_{2.5}$ for Summer (07/20/2015)

(**b**)
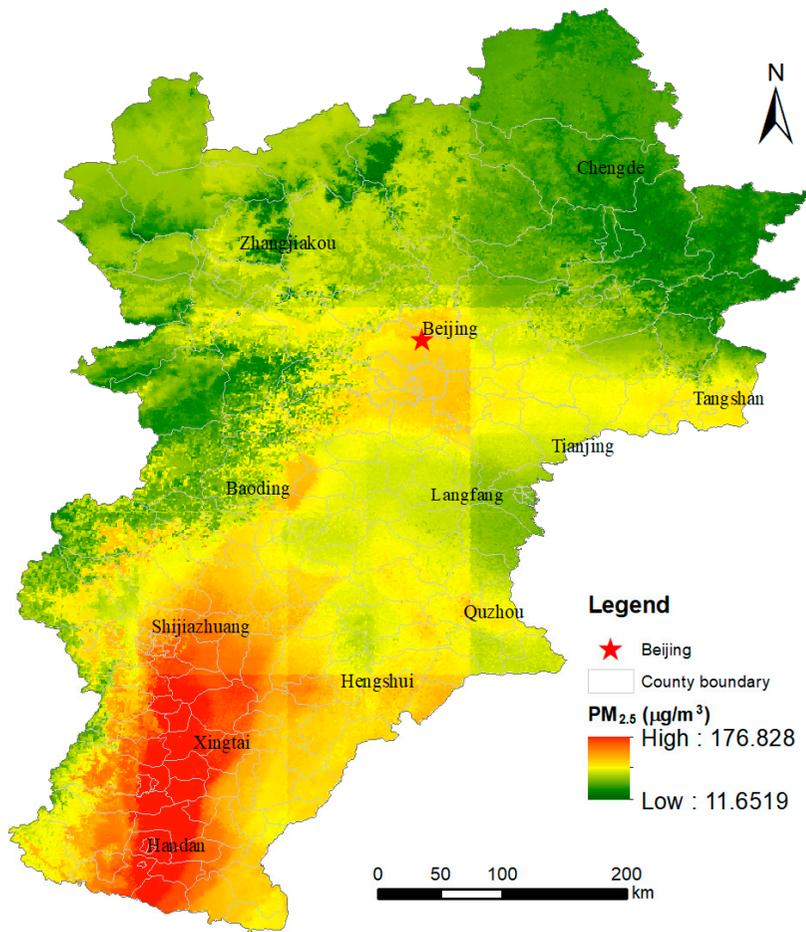
c. PM₂.₅ for Autumn (10/20/2015)

(c)
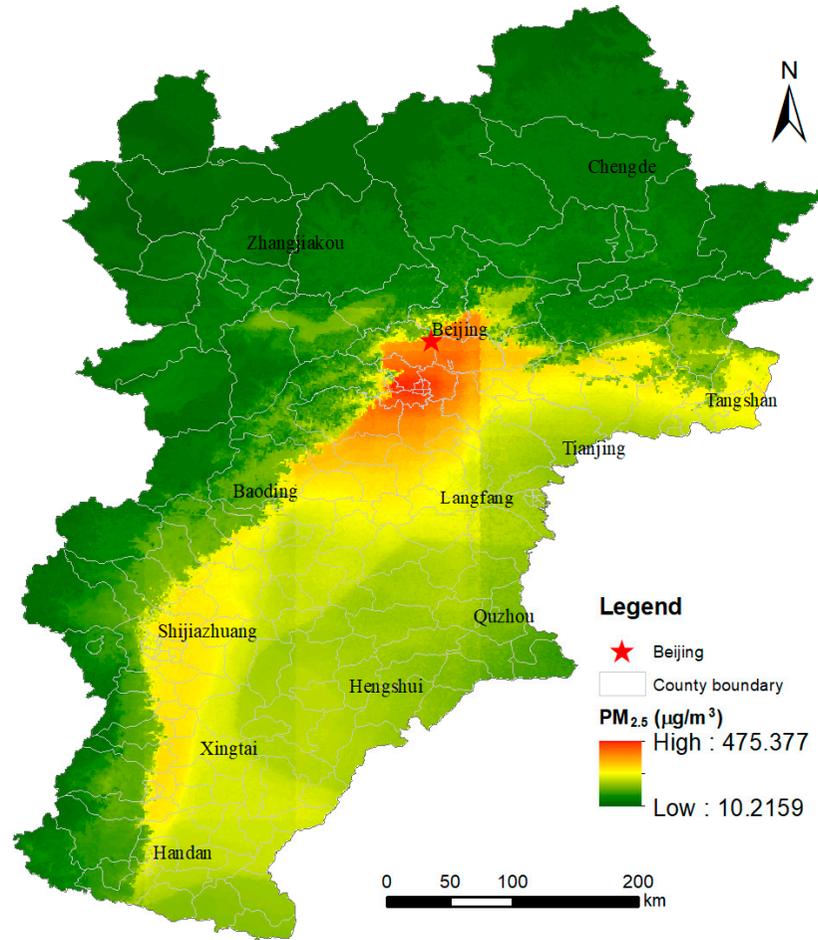
d. PM₂.₅ for Winter (12/01/2015)

(d)

Figure S11. Grid surfaces of predicted PM$_{2.5}$ (μg/m^3) by XGBoost for four seasonal days (a: spring; b: summer; c: autumn; d: winter) in 2015 (representation with spatial discontinuity).
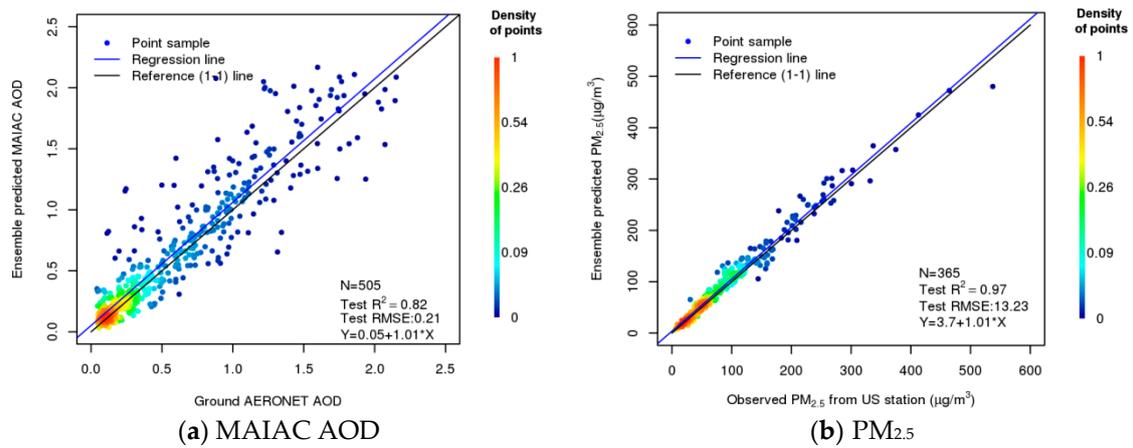
(**a**) MAIAC AOD   (**b**) PM2.5

Figure S12. Scatterplots between observed vs. predicted values in the independent tests of (**a**) MAIAC AOD using two AERONET sites and (**b**) PM2.5 using the US embassy monitoring station.