A national-scale 1 km resolution $PM_{2.5}$ estimation model over Japan using MAIAC AOD and a two-stage random forest model

Chau-Ren Jung[a,b], Wei-Ting Chen[c], and Shoji F. Nakayama[a,*]

[a]Japan Environment and Children's Study Programme Office, National Institute for Environmental Studies, Tsukuba, Japan

[b]Department of Public Health, College of Public Health, China Medical University, Taichung, Taiwan

[c]Department of Atmospheric Sciences, National Taiwan University, Taipei, Taiwan

Correspondence to:

Dr Shoji F. Nakayama, MD, PhD

Japan Environment and Children's Study Programme Office, National Institute for Environmental Studies

16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan

Telephone: +81 (29) 850-2786

E-mail: fabre@nies.go.jp

Table S1. Summary of data sources, coverage years, and spatial and temporal resolutions.

Variable	Source	Year	Spatial resolution	Temporal resolution
Ground $PM_{2.5}$ measurements	National Institute for Environmental Studies	2011–2016	Points	Daily
Multi-Angle Implementation of Atmospheric Correction (MAIAC) Aerosol Optical Depth	NASA	2011–2016	1 km	Daily (one from Terra, which overpasses at 10:30 a.m., and one from Aqua, which overpasses at 1:30 p.m.)
Meteorological variables				
Temperature	Japan Meteorological Agency	2011–2016	1 km (using regression-kriging)	Daily
Relative humidity	Japan Meteorological Agency	2011–2016	1 km (using ordinary kriging)	Daily
Precipitation	Japan Meteorological Agency	2011–2016	1 km (using ordinary kriging)	Daily
Surface pressure	Japan Meteorological Agency	2011–2016	1 km (using regression kriging)	Daily
10 m height zonal wind (u10)	European Centre for Medium-Range Weather	2011–2016	0.125° (approximately 13 km)	Daily

	Forecasts (ECMWF)			
10 m height meridional wind (v10)	European Centre for Medium-Range Weather Forecasts (ECMWF)	2011–2016	0.125° (approximately 13 km)	Daily
Boundary layer height	European Centre for Medium-Range Weather Forecasts (ECMWF)	2011–2016	0.125° (approximately 13 km)	Daily
Cloud fraction	NASA	2011–2016	5 km	Daily
Land use data				
Urban and built-up areas	Japan Aerospace Exploration Agency	2006–2011	50 m	None
Industrial areas	Japan Ministry of Land, Infrastructure, Transport and Tourism	2009	250 m	None
Road networks	Geoinformation Authority of Japan	2016	None (line data)	None
Normalized difference vegetation index (NDVI)	NASA	2011–2016	250 m	16 days
Population count	WorldPop	2011–2016	100 m	Annual
Elevation	NASA	GDM v2 announced in 2011	30 m	None

Table S2. Descriptive statistics of 16 predictors from 2011 to 2016 (mean ± standard deviation).

Year	Temp (°C)	Hum (%)	CF (%)	BLH (m)	Prec (mm)	SP (hPa)	u10 (m/s)	v10 (m/s)
2011	10.71±10.29	71.77±10.85	0.74±0.33	700.27±341.09	4.59±13.40	974.20±41.25	1.23±2.78	-0.27±2.75
2012	10.54±10.53	72.78±10.31	0.74±0.33	708.75±344.33	4.29±11.01	974.00±41.25	0.88±2.87	-0.34±2.68
2013	10.94±10.24	72.29±10.78	0.71±0.35	709.54±335.94	4.29±11.49	973.40±41.28	1.31±2.78	-0.34±2.74
2014	10.65±9.94	72.37±11.31	0.72±0.33	692.43±300.37	4.27±11.50	974.00±41.53	1.11±2.81	-0.42±2.75
2015	11.38±9.27	73.68±11.37	0.71±0.35	659.97±312.76	4.32±10.66	974.50±41.21	1.01±2.80	-0.33±2.69
2016	11.42±9.93	73.91±10.97	0.73±0.34	669.00±306.51	4.51±11.05	975.20±40.59	0.99±2.70	-0.26±2.66

Year	NDVI (unitless)	Industril area (m^2)	Urban area (m^2)	Elevation (m)
2011	0.51±0.33	4954.00±47758.17	73510.00±229570.00	372.60±391.02
2012	0.51±0.34			
2013	0.52±0.34			
2014	0.52±0.33			
2015	0.54±0.34			
2016	0.54±0.34			

Abbreviations: BLH, boundary layer height; CF, cloud fraction; Hum, relative humidity; NDVI, normalized difference vegetation index; Prec, precipitation; SP, surface pressure; Temp, temperature; u10, 10 m height zonal wind; v10, 10 m height meridional wind.

Table S2. (continued).

Year	Road length (m)	Distoprim (km)	Distohigh (km)	Pop (number/10,000 m^2)
2011	335.20±566.95	4.13±5.80	16.31±23.08	2.26±8.80
2012				2.26±8.95
2013				2.27±9.05
2014				2.27±9.03
2015				2.27±9.05
2016				2.27±9.26

Abbreviations: Distohigh, distance to the nearest highway; Distoprim, distance to the nearest primary road; Pop, population count.

Table S3. Descriptive statistics for the Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol optical depth (AOD) and AErosol RObotic NETwork (AERONET) AOD.

Variable	Mean±SD	Median	Minimum	Q1	Q3	Maximum	IQR
MAIAC AOD	0.196±0.132	0.169	0.019	0.109	0.254	1.693	0.145
AERONET AOD	0.202±0.133	0.167	0.018	0.110	0.261	1.295	0.151

Abbreviations: SD, standard deviation; Q1, 25^{th} percentile; Q3, 75^{th} percentile; IQR, interquartile range.

Table S4. Comparison of model performances between this and other studies that applied machine learning algorithms to develop satellite-based $PM_{2.5}$ models.

No.	Author (published year)	Study area	Study period	AOD product	Model	Model performance CV R^2 of model (RMSE)
1	This study	Japan	2011–2016	1 km Terra and Aqua MAIAC AOD (MCD19A2)	Random forest	10-fold CV R^2 of 0.86 (3.02 μg/m^3)
2	Di et al. (2016)	USA	2000–2012	1 km Terra and Aqua MAIAC AOD	Deep learning	10-fold CV R^2 of 0.84 (2.94 μg/m^3)
3	Hu et al. (2017)	USA	2011	10 km Aqua collection 6 L2 DT AOD (MYD04_L2)	Random forest	10-fold CV R^2 of 0.80 (2.83 μg/m^3)
4	Brokamp et al. (2018)	Cincinnati, OH, USA	2000–2015	3 km Terra and Aqua collection 6 L2 DT AOD (MOD04_L2 and MYD04_L2)	Random forest	Leave-one-out CV R^2 of 0.91 (2.22 μg/m^3)
5	Chen et al. (2018)	Mainland China	2005–2016	10 km Aqua collection 6 L2 DT and DB AOD	Random forest	10-fold CV R^2 of 0.83 (28.1 μg/m^3)
6	Huang et al. (2018)	North China Plain area, China	2013–2016	1 km Terra and Aqua MAIAC AOD	Random forest	10-fold CV R^2 of 0.88 (14.89 μg/m^3)*
7	Zhang et al. (2018)	Sichuan Basin, China	2013–2015	3 km Terra and Aqua collection L2 DT AOD	Random forest	10-fold CV R^2 of 0.86 (15.9 μg/m^3)
8	Chen et al. (2019)	Mainland China	2014–2015	3 km Terra and Aqua collection 6 L2 DT AOD and 10 km DB AOD	Non-linear exposure-lag-response model with XGBoost	10-fold CV R^2 of 0.86 (14.98 μg/m^3)
9	Stafoggia et al. (2019)	Italy	2013–2015	1 km Aqua MAIAC AOD	Random forest	10-fold CV R^2 of 0.79, 0.78 and 0.81 for 2013, 2014 and

						2015, respectively (6.59, 5.36 and 6.39 μg/m^3, respectively)
10	Wang and Sun (2019)	Beijing-Tianjin-Hebei, China	2014	10 km Aqua collection 6 DB AOD	Deep learning	10-fold CV R^2 of 0.87 (27.11 μg/m^3)
11	Yang et al. (2019)	Fuzhou, China	2014–2016	3 km Terra collection 6 L2 DT AOD	Linear mixed effect model with support vector machine	10-fold CV R^2 of 0.77 (9.51 μg/m^3)
12	Joharestani et al. (2019)	Tehran, Iran	2015–2018	10 and 3 km Terra collection 6 L2 DT AOD (MOD04_L2 and MOD04_3k)	Random forest, XGBoost and deep learning	10-fold CV R^2 of 0.66 (15.30 μg/m^3), 0.67 (15.15 μg/m^3) and 0.63 (15.89 μg/m^3) for random forest, XGBoost and deep learning model, respectively (after including 3 km AOD product in models)
12	Chen et al. (2020)	Guangdong–Hong Kong–Macao Greater Bay Area, China	2016–2018	1 km Terra and Aqua MAIAC AOD (MCD19A2)	Random forest	10-fold CV R^2 of 0.937, 0.905 and 0.884 for 2016, 2017 and 2018, respectively (3.527, 3.78 and 3.633 μg/m^3, respectively)
13	Schneider et al. (2020)	Great Britain, UK	2008–2018	1 km Terra and Aqua MAIAC AOD (MCD19A2)	Random forest	10-fold CV R^2 of 0.767 (4.042 μg/m^3)
14	Shtein et al. (2020)	Italy	2013–2015	1 km Aqua MAIAC AOD	Ensemble model combining a linear mixed effect model, random forest and XGBoost	CV R^2 of 0.79, 0.79 and 0.81 for 2013, 2014 and 2015, respectively (6.56, 5.29 and 6.34 μg/m^3, respectively)
15	Zhang et al. (2021)	Mainland China	2017	3 km Terra and Aqua collection 6 L2 DT	Gradient boosting	10-fold CV R^2 of 0.81 (11.57 μg/m^3)

				AOD (MOD04_3k and MYD04_3k)		

Note: *for estimating monthly average $PM_{2.5}$ concentrations.

Abbreviations: AOD, aerosol optical depth; CV, cross-validation; MAIAC, Multi-Angle Implementation of Atmospheric Correction; R^2, coefficient of determination; RMSE, root mean square error; XGBoost, extreme gradient boosting.

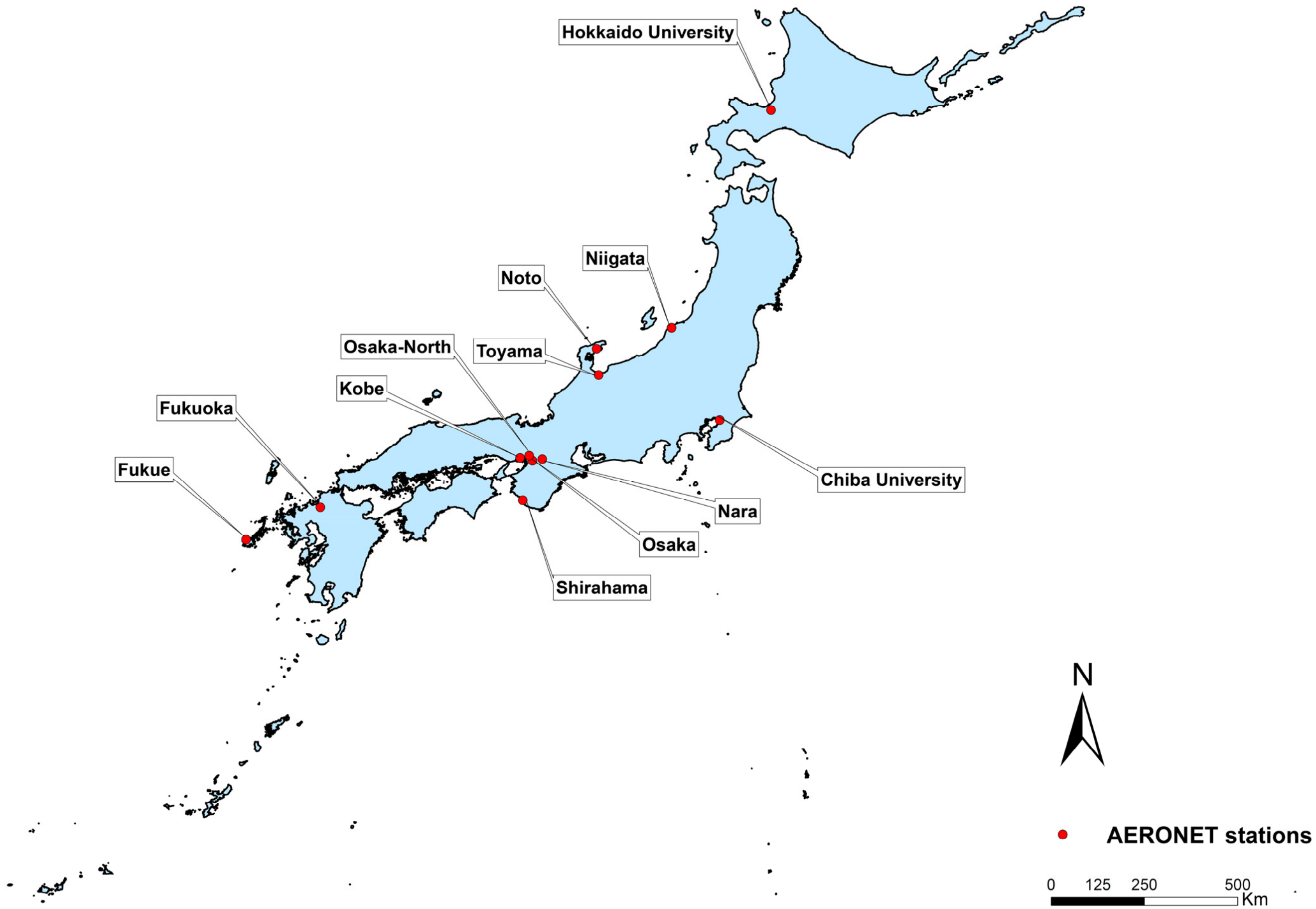

Figure S1. Locations of 12 AErosol RObotic NETwork (AERONET) stations in Japan (Hokkaido University, Niigata, Noto, Toyama, Chiba University, Osaka-North, Nara, Osaka, Kobe, Shirahama, Fukuoka and Fukue).

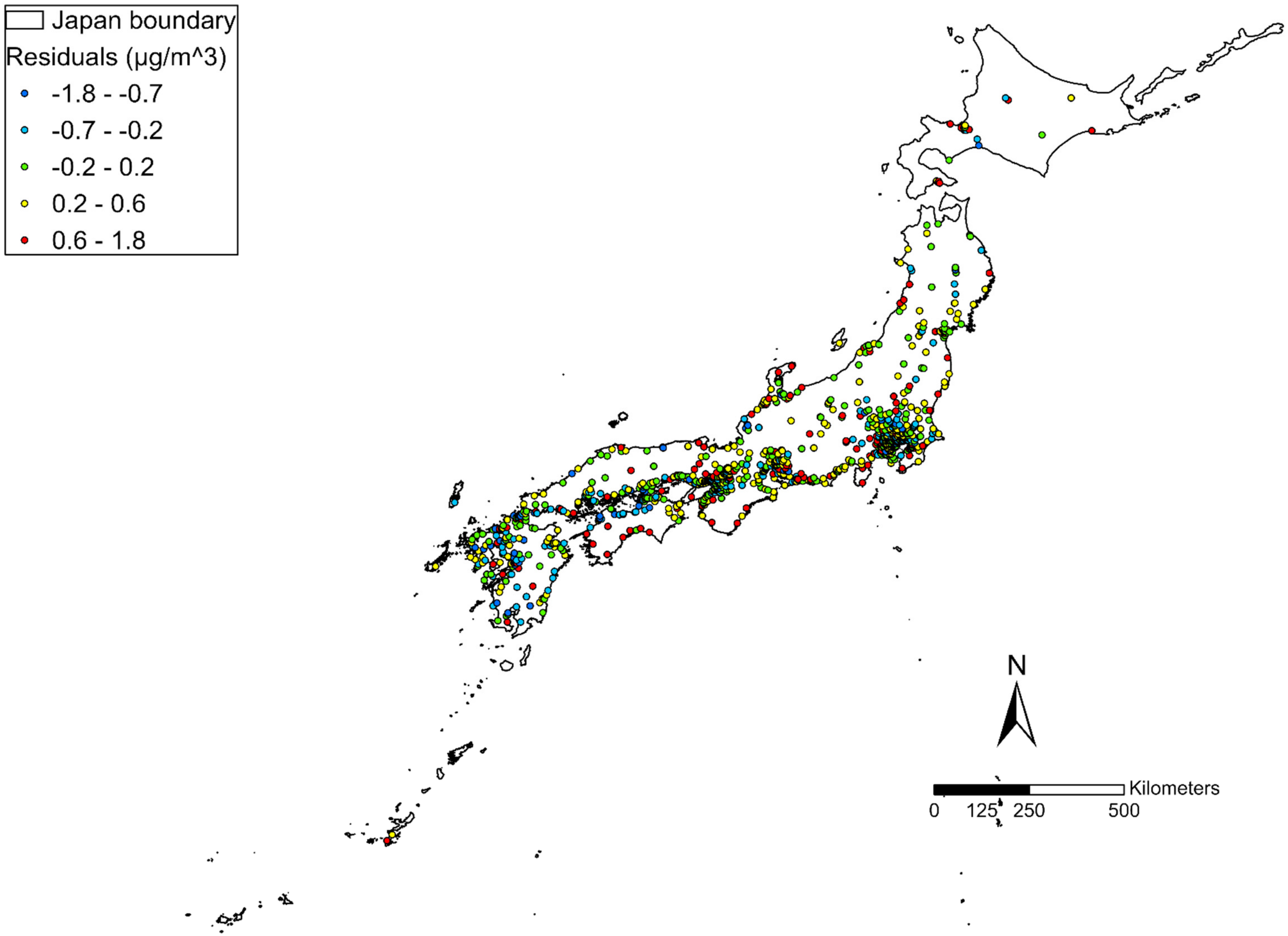

Figure S2. The spatial distribution of average residuals (differences between estimated $PM_{2.5}$ and in situ measurements) based on the random forest model.

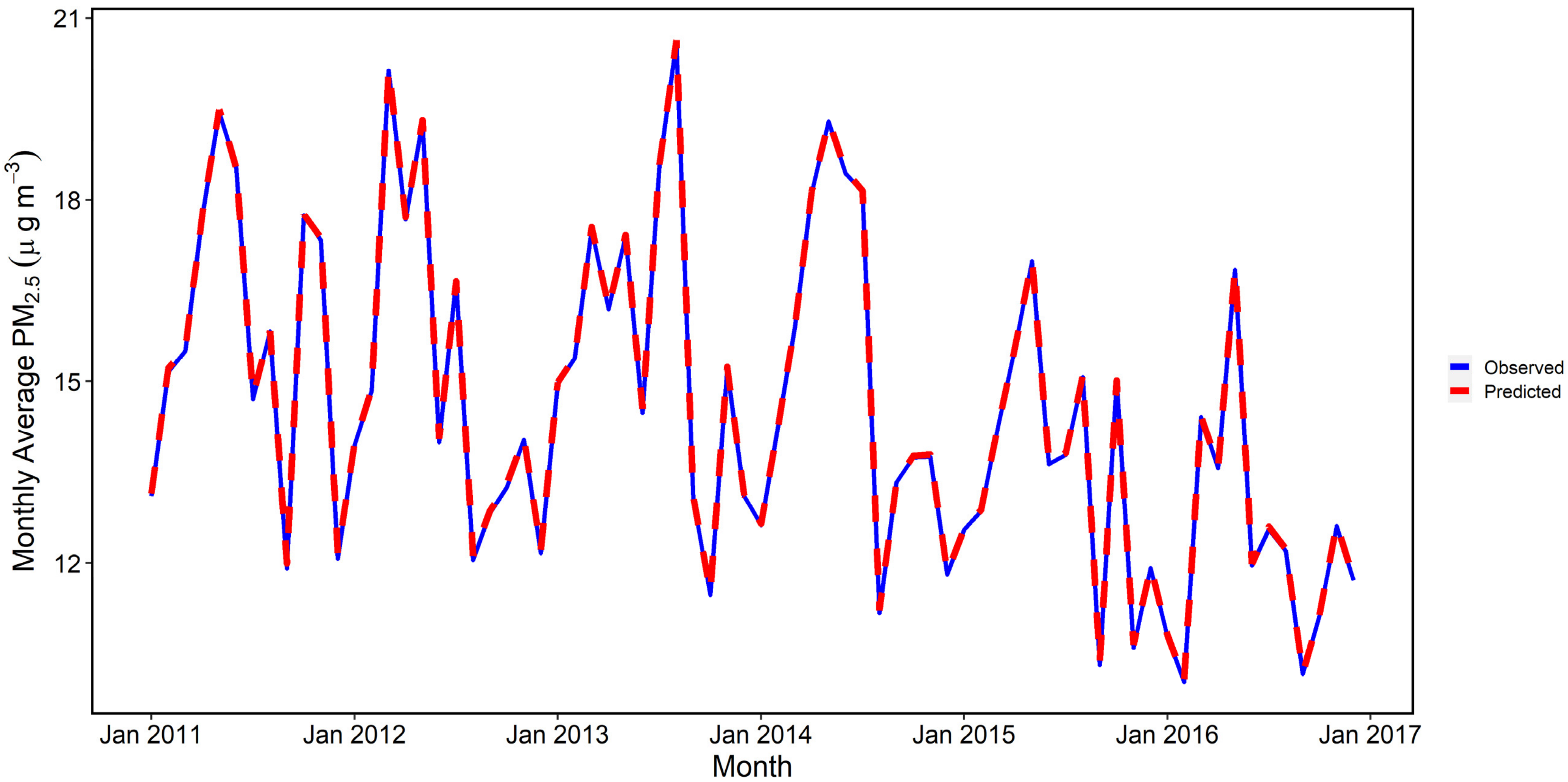

Figure S3. Time series variation of monthly average observed and predicted $PM_{2.5}$ concentrations during 2011–2016.

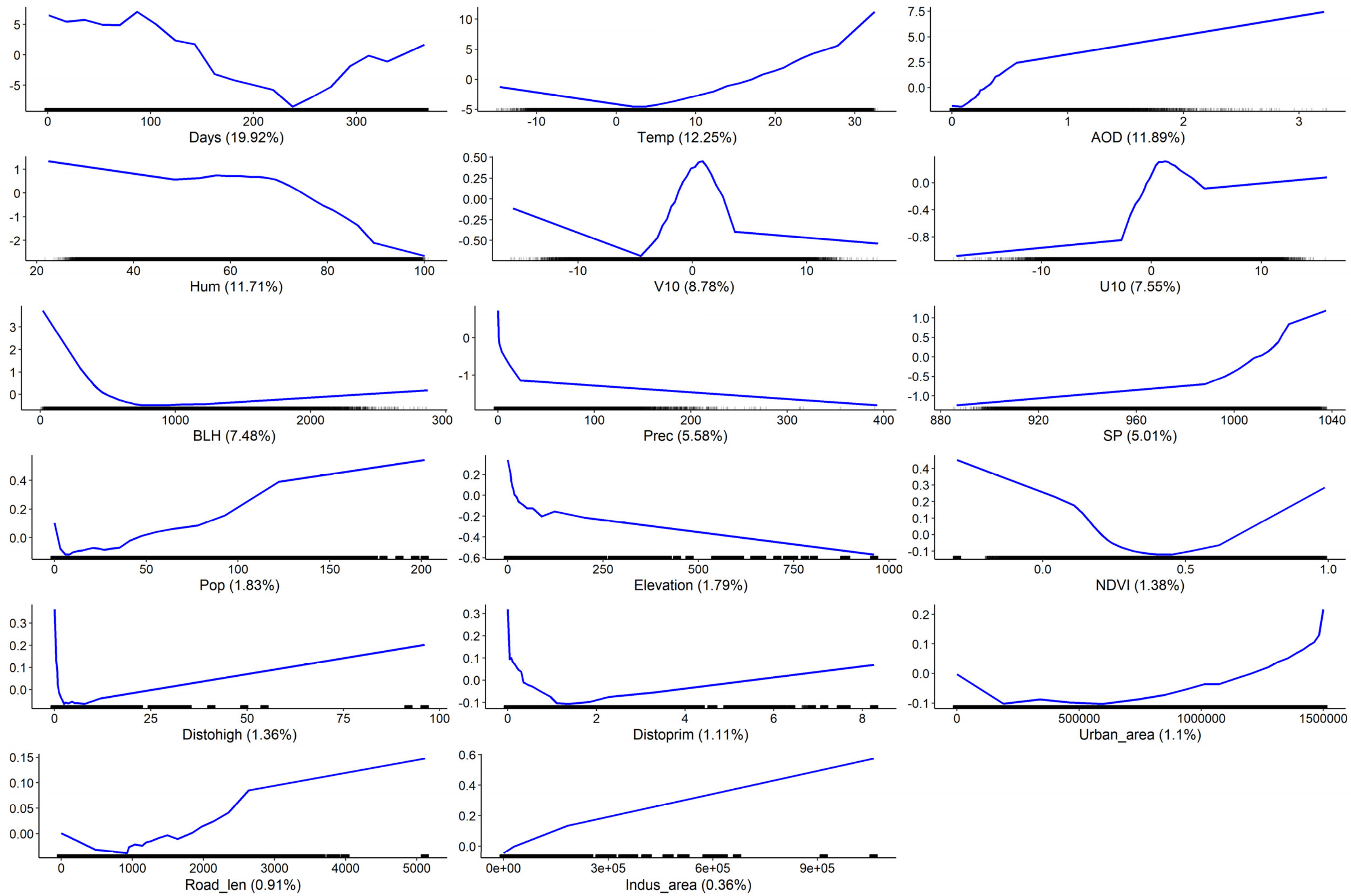

Figure S4. Accumulated local effect plots for the effects of predictors on estimated $PM_{2.5}$ values. Numbers in brackets show the relative permutation importance (%).