Full-coverage 1 km Daily Ambient $PM_{2.5}$ and O_3 Concentrations of China in 2005-2017 Based on Multi-variable Random Forest Model

Supplementary materials

Runmei Ma[1#], Jie Ban[1#], Qing Wang[1#], Yayi Zhang[1], Yang Yang[2], Shenshen Li[3], Wenjiao Shi[4, 5], Tiantian Li[1*]

[1] China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing, 100021, China

[2] Institute of Urban Meteorology, China Meteorological Administration, Beijing, 100089, China

[3] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, 100101, China

[4] Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

[5] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

[#] Runmei Ma, Jie Ban and Qing Wang are the co-first authors.

[*] Corresponding author: E-mail: litiantian@nieh.chinacdc.cn

Content

Table S1 The basic information of model variables

Category	Classification	Spatial Resolution	Temporal Resolution	Time Frame	Source
$PM_{2.5}$ measurements	$PM_{2.5}$ measurements	-	Day	2013/01/18-2017/12/31	CEMC (http://www.cnemc.cn/)
O_3 measurements	O_3 measurements	-	Day	2013/01/18-2017/12/31	CEMC (http://www.cnemc.cn/)
Meteorological data	Boundary layer height, Surface pressure, 2 meter dewpoint temperature, Evaporation, Albedo, Low cloud cover, Medium cloud cover, High cloud cover, Total precipitation, 10 meter U wind component, 10 meter V wind component, 2 meter temperature, Surface solar radiation downwards	0.125°	6/12 hours	2005-2017	ERA-Interim (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/)
Satellite derived data	Aerosol optical depth	3 km	Day	2005-2017	NASA-Laads (https://ladsweb.modaps.eosdis.nasa.gov/)
GEOS-Chem	Modeling O_3 from GEOS-Chem model outputs	2°×2.5°	2 hours	2005-2016	GEOS-Chem model output (Ma *et al.*, 2021)
Socio-economic variables	Population, GDP	1 km	Year	2005、2010	RESDC (http://www.resdc.cn)
	Elevation	1 km	Year	2010	RESDC (http://www.resdc.cn)
Land-Use Variables	Road (railway, high speed road, provincial road, county road, town road)	Vector	Year	2016	RESDC (http://www.resdc.cn)
	Land use	300m	Year	2005-2017	UCL-Geomatics (http://maps.elie.ucl.ac.be/CCI/viewer/）
	NDVI	300 m	Month	2005-2016	Geospatial Data Cloud (https://www.gscloud.cn/)
Dummy Variables	Season, Providence, Month, Year	- -	- -	- -	-

Table S2 The results of descriptive analysis for variables included in $PM_{2.5}$ and O_3-8hamx model construction (N = 197,366)

Variables	Minimum	Median	Mean	Maximum	SD
AOD	0.00	0.55	0.68	5.52	0.53
GEOS-Chem	0.00	19.02	20.36	82.80	20.03
Albedo	0.07	0.16	0.17	0.29	0.03
Boundary layer height (m)	9.47	265.32	308.74	3234.18	250.45
2 meter dewpoint temperature (K)	232.07	283.40	281.51	301.14	12.49
Evaporation(m)	-0.01	0.00	0.00	0.00	0.00
High cloud cover	0.00	0.21	0.30	1.00	0.30
Medium cloud cover	0.00	0.05	0.18	1.00	0.25
Low cloud cover	0.00	0.12	0.21	1.00	0.24
Surface pressure (Pa)	54485	98961	95799	104525	7746
Surface solar radiation downwards (w/(m^2·s))	5069412	5593930	5571471	11177510	2431341
2 meter temperature (℃)	-37.67	16.70	14.72	36.37	10.96
Total precipitation(m)	0.00	0.00	0.00	0.22	0.01
10 meter U wind component (m/s)	-13.99	-0.23	-0.11	11.70	1.83
10 meter V wind component (m/s)	-14.81	-0.05	-0.13	11.96	2.30
GDP (RMB Yuan)	0.00	18757.80	24567.45	146901.00	23141.48
DEM (m)	-10	75	353.67	4509	621.08
Population (persons)	0	6301	6746	51338	6042
urban coverage rate (0~1)	0	0.78	0.66	1.00	0.31
NDVI (0-1)	0	0.41	0.42	1.00	0.20
Country road (m)	0	0	503.33	5914.65	980.92
highway (m)	0	0	16.11	5091.96	207.82
provincial road (m)	0	0	531.50	6948.42	1009.39
railway (m)	0	0	5.82	1104.12	73.85
Town road (m)	0	4717.66	4809.80	15582.40	3058.31

Table S3 The model performance of $PM_{2.5}$ and O_3-8hamx in different provinces

Provinces	$PM_{2.5}$			O_3-8hamx		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Shanghai	0.94	9.14	5.69	0.84	19.41	12.62
Beijing	0.92	17.77	10.84	0.92	17.63	11.51
Hubei	0.91	13.00	8.35	0.76	24.62	16.37
Hebei	0.91	21.05	12.81	0.81	25.15	16.11
Sichuan	0.9	12.24	8.38	0.74	22.59	15.89
Jiangsu	0.89	12.63	7.69	0.75	25.74	16.19
Hunan	0.89	12.32	8.09	0.74	21.88	15
Jilin	0.88	15.80	9.53	0.78	18.73	13.43
Tianjing	0.88	19.59	11.93	0.83	22.52	15.41
Guangxi	0.88	10.29	6.53	0.68	21.64	14.3
Henan	0.88	19.13	11.54	0.83	22.26	15.2
Chongqing	0.88	12.70	8.54	0.86	20.31	14.39
Zhejiang	0.87	12.44	7.62	0.74	24.75	16.84
Liaoning	0.87	15.91	9.76	0.79	22.19	15.04
Heilongjiang	0.86	18.79	10.71	0.6	21.92	14.66
Hainan	0.85	6.10	4.21	0.71	18.22	13.87
Jiangxi	0.84	11.00	7.32	0.77	19.94	13.47
Guizhou	0.82	10.24	7.19	0.62	19.42	13.88
Shandong	0.81	23.04	11.38	0.79	25.01	16.73
Guangdong	0.79	11.45	5.86	0.78	23.04	14.97
Shanxi	0.77	22.77	12.7	0.81	21.57	15.08
Shanxi	0.75	31.17	12.65	0.78	21.57	15.08
Xinjiang	0.74	38.70	19.08	0.66	23.76	15.75
Fujian	0.73	8.56	5.61	0.68	20.16	14.69
Inner Mongolia	0.72	18.29	9.71	0.71	22.92	14.1
Ningxia	0.7	17.46	11.59	0.77	19.56	14.27
Yunnan	0.69	9.50	6.5	0.67	18.58	13.53
Anhui	0.69	23.77	9.69	0.6	30.4	21.32
Gansu	0.68	17.31	10.75	0.55	25.15	15.53
Qinghai	0.62	16.47	11.22	0.66	22.12	15.4
Xizang	0.6	11.18	7.22	0.62	15.57	10.15
Taiwan	-	-	-	-	-	-
Aomen	-	-	-	-	-	-
Hongkong	-	-	-	-	-	-

Table S4-1. Feature importance of each variable in the $PM_{2.5}$ model

Variable	Value	Variable	Value
Boundary layer height lag1	0.172	2 meter temperature lag2	0.013
AOD	0.056	Total precipitation	0.012
DEM	0.049	Total precipitation lag1	0.012
Boundary layer height	0.043	Evaporationlag2	0.011
Evaporation lag1	0.038	High cloud cover lag2	0.011
Season	0.037	Surface pressure	0.011
2 meter dewpoint temperature	0.036	Medium cloud cover	0.011
2 meter dewpoint temperature lag2	0.026	2 meter temperature lag1	0.011
AOD lag1	0.023	High cloud cover lag1	0.010
Boundary layer height lag2	0.021	Medium cloud cover lag1	0.009
10 meter V wind component lag2	0.020	Medium cloud cover lag2	0.009
10 meter V wind component lag1	0.020	Low cloud cover	0.009
10 meter U wind component_lag1	0.020	Surface pressure lag1	0.009
Province	0.020	Low cloud cover lag1	0.008
Year	0.019	Low cloud cover lag2	0.007
10 meter U wind component_lag2	0.018	GDP	0.007
Surface solar radiation downwards	0.018	Population	0.007
Evaporation	0.017	NDVI	0.006
Surface solar radiation downwards lag1	0.015	Albedo	0.006
2 meter dewpoint temperature lag1	0.015	Total precipitation lag2	0.006
Surface pressure lag2	0.014	Albedo lag2	0.005
Surface solar radiation downwards lag2	0.014	Albedo lag1	0.004
10 meter V wind component	0.014	Town road	0.004
High cloud cover	0.014	Country road	0.004
AOD lag2	0.014	Land use	0.003
Month	0.013	Providence road	0.002
10 meter U wind component	0.013	Railway	0.0001
2 meter temperature	0.013	High speed road	0.0001

Table S4-2. Feature importance of each variable in the O_3-8hmax model

Variable	Value	Variable	Value
Surface solar radiation downwards	34.18%	Surface pressure lag2	1.07%
Outputs from GEOS-Chem model	7.24%	10 meter U wind component lag2	1.05%
2 meter temperature	4.22%	2 meter dewpoint temperature lag1	1.05%
Surface solar radiation downwards lag1	3.89%	High cloud cover lag2	1.04%
Year	2.17%	Population	1.01%
Dem	1.88%	Town road	1.00%
Boundary layer height	1.86%	Medium cloud cover lag1	0.99%
10 meter V wind component	1.80%	2 meter temperature lag1	0.97%
Low cloud cover	1.66%	NDVI	0.97%
Province	1.56%	Low cloud cover lag2	0.96%
2 meter dewpoint temperature	1.44%	2 meter temperature lag2	0.94%
10 meter V wind component lag1	1.35%	Evaporation lag2	0.93%
10 meter U wind component	1.34%	Evaporation lag1	0.93%
Evaporation	1.33%	Total precipitation	0.92%
Low cloud cover lag1	1.24%	Medium cloud cover lag2	0.91%
10 meter U wind component lag1	1.20%	Surface pressure lag1	0.87%
GDP	1.18%	Total precipitation lag2	0.77%
Boundary layer height lag1	1.17%	Month	0.76%
Total precipitation lag1	1.12%	Land use	0.44%
Boundary layer height lag2	1.12%	County road	0.41%
2 meter dewpoint temperature lag2	1.11%	Province road	0.39%
Surface solar radiation downwards lag2	1.10%	Albedo	0.32%
High cloud cover	1.09%	Albedo lag2	0.31%
Medium cloud cover	1.09%	Albedo lag1	0.22%
High cloud cover lag1	1.08%	Season	0.17%
Surface pressure	1.08%	Railway	0.01%
10 meter V wind component lag2	1.08%	High speed road	0.01%

Table S5 The real and simulated concentration of $PM_{2.5}$ and O_3-8hamx in different provinces

Provinces	$PM_{2.5}$		O_3-8hamx	
	Real concentration (μg/m^3)	Simulated concentration (μg/m^3)	Real concentration (μg/m^3)	Simulated concentration (μg/m^3)
Shanghai	49.62	49.66	104.93	106.61
Beijing	72.49	72.48	92.27	92.96
Hubei	62.71	62.92	88.92	88.82
Hebei	80.63	80.69	90.79	91.61
Sichuan	49.88	49.80	80.16	80.10
Jiangsu	57.47	57.20	97.70	98.62
Hunan	53.95	53.83	80.89	80.98
Jilin	50.18	50.00	85.90	87.54
Tianjing	77.02	77.47	84.95	83.09
Guangxi	42.04	42.17	81.09	80.33
Henan	75.16	75.18	94.39	94.88
Chongqing	56.33	55.71	69.68	68.56
Zhejiang	48.56	48.45	90.86	92.84
Liaoning	51.60	51.55	89.85	89.88
Heilongjiang	43.91	43.54	70.37	69.80
Hainan	20.60	19.86	73.34	69.47
Jiangxi	45.53	45.11	80.33	79.18
Guizhou	37.08	36.63	71.78	70.01
Shandong	65.62	65.73	98.84	100.08
Guangdong	36.74	36.56	88.58	88.84
Shanxi	60.26	60.36	85.57	84.96
Shanxi	63.95	63.78	82.73	82.56
Xinjiang	59.86	59.24	78.05	78.28
Fujian	28.55	27.89	79.99	79.56
Inner Mongolia	40.11	39.76	87.39	86.73
Ningxia	45.63	44.85	89.46	88.60
Yunnan	27.29	27.04	77.47	76.68
Anhui	58.16	57.96	83.76	81.80
Gansu	41.25	40.70	91.43	90.58
Qinghai	43.33	43.42	87.25	87.94
Xizang	23.83	23.40	91.62	91.26
Taiwan	-	-	-	-
Aomen	-	-	-	-
Hongkong	-	-	-	-

Table S6-1 The characteristic of similar modeling work for $PM_{2.5}$

Model	Location	Time period	Spatial Resolution	Temporal Resolution	Model Validation			Predictive power	
					R^2	RMSE	MAE	Daily	Monthly
RF (This study)	China	2005-2017	1km	Daily	0.85	17.72	9.37	0.49	0.65
RF(Chen *et al.*, 2018)	China	2005-2016	10km	Daily	0.83	18.08	-	-	-
Geoi-DBN(Li *et al.*, 2017)	China	2015	10km	Daily	0.94	13.68			
GWR (Ma *et al.*, 2014)	China	2012-2013	10km	Daily	0.64	32.98	21.25	-	-
GWR(Hammer *et al.*, 2020)	Global	1998-2018	1km	Annual	0.90-0.92				
STET(Wei *et al.*, 2020)	China	2018	1km	Daily	0.89	10.35	6.71	0.6	0.8
Gaussian process models(Yu *et al.*, 2017)	China	2013	10km	Daily	0.81	21.87			
STET(Wei *et al.*, 2021)	China	2013-2018	1km	Daily	0.86-0.90	10-18.4	6.4-11.5		0.80
TSAM(Fang *et al.*, 2016)	China	2013-2014	10km	Daily	0.8	22.75	15.99	-	-
Two-stage(Ma *et al.*, 2016)	China	2004-2013	10km	Daily	0.79	27.42			
ML+GAM(Xue *et al.*, 2019)	China	2000-2016	10km	Daily	0.55	30.2	18.9		
Two-stage(Geng *et al.*, 2021)	China	Near Real-Time	10km	Daily	0.80−0.88	13.9−22.1			
XGBoost(Chen *et al.*, 2019)	China	2014-2015	3km	Daily	0.86	14.98			
STRF(Wei *et al.*, 2019)	China	2016	1km	Daily	0.85	15.57			

Table S6-2 The characteristic of similar modeling work for O_3

Model	location	Resolution		R^2	RMSE
		Temporal	Spatial		
RF (This study)	China	O_3-8hmax	1km×1km	0.77	21.76
XGBoost(Liu *et al.*, 2020)	China	O3-8hmax	0.1*0.1°	0.78	21.47
RF(Zhan *et al.*, 2018)	China	O_3-8hmax	0.1*0.1°	0.69	26
STE-ResNet(Li and Cheng, 2021)	Greater Bay Area, China	O_3-8hmax	0.05*0.05°	0.93	12.99
GWR(Zhang *et al.*, 2020)	Eastern China	Monthly average	0.25*0.25°	0.81	
CNN(Di *et al.*, 2017)	America	O_3-8hmax	1km×1km	0.76	7.33ppb
LUR(Wang *et al.*, 2015)	Six U. S. metropolitan regions	Annual average	50m	0.65-0.88	
LUR(Malmqvist *et al.*, 2014)	Two Swedish cities	Weekly average; daily average	50m	0.40, 0.67	
LUR+Bayes(Adam-Poupart *et al.*, 2014)	Quebec, Canada	O_3-8hmax	1 km×1 km	0.653	7.06
LUR(Kerckhoffs *et al.*, 2015)	Netherlands	Summer average and annual average	50×50m	0.71;0.77	
LUR(Beelen *et al.*, 2009)	Europe	Annual average	1×1 km	0.53	
CTM+LUR(Wang *et al.*, 2016)	Los Angeles Basin	Two weeks average	4×4 km	0.84	3.62
LUR(Wolf *et al.*, 2017)	Augsburg, Germany	Annual average	1km	0.92	
LUR(Son *et al.*, 2018)	Mexico	Hourly average, monthly average, semiannual average, annual average	30*30m	0.65	16.77

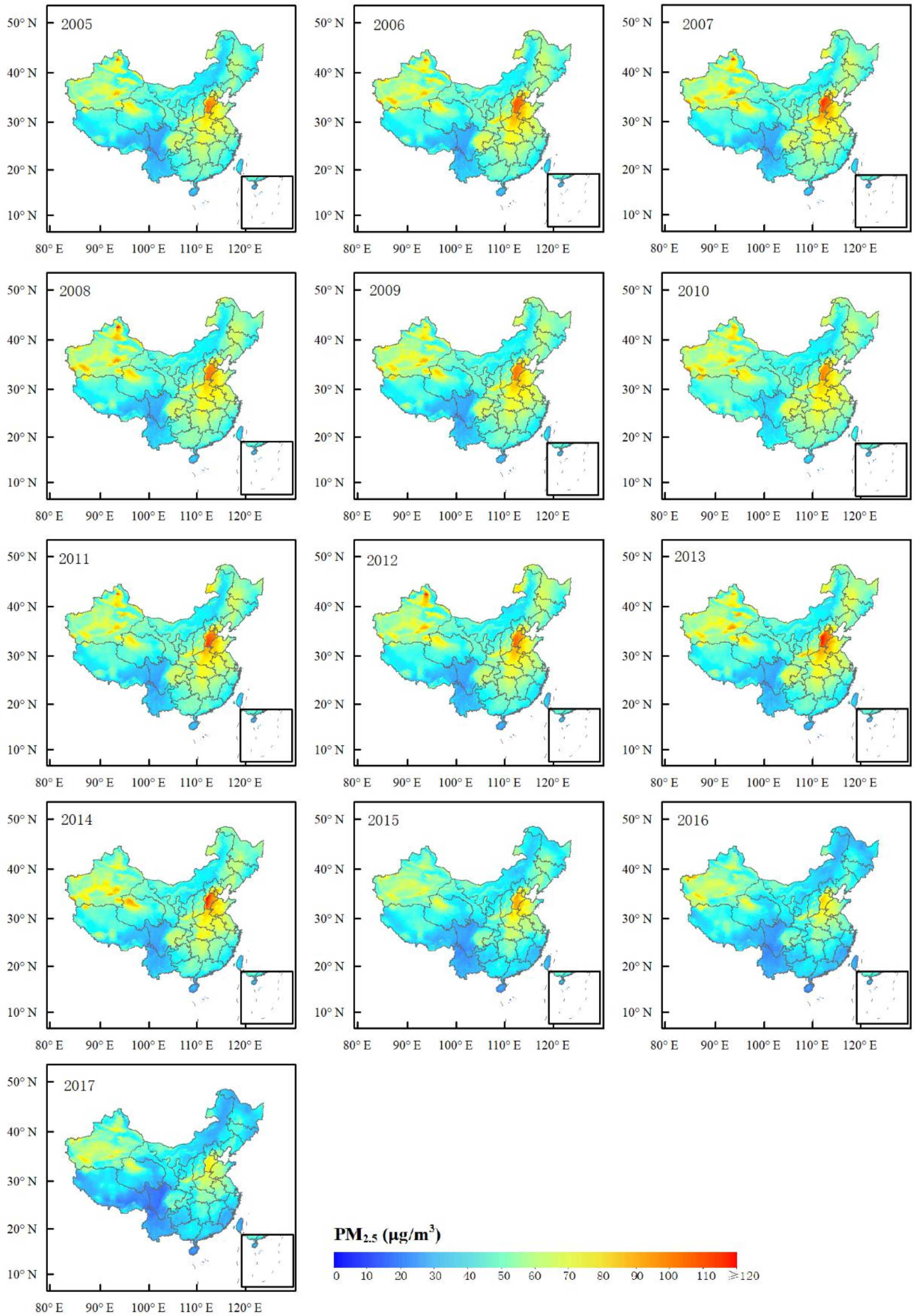

Fig. S1 Simulated annual mean PM2.5 concentration based on random forest model in China from 2005 to 2017

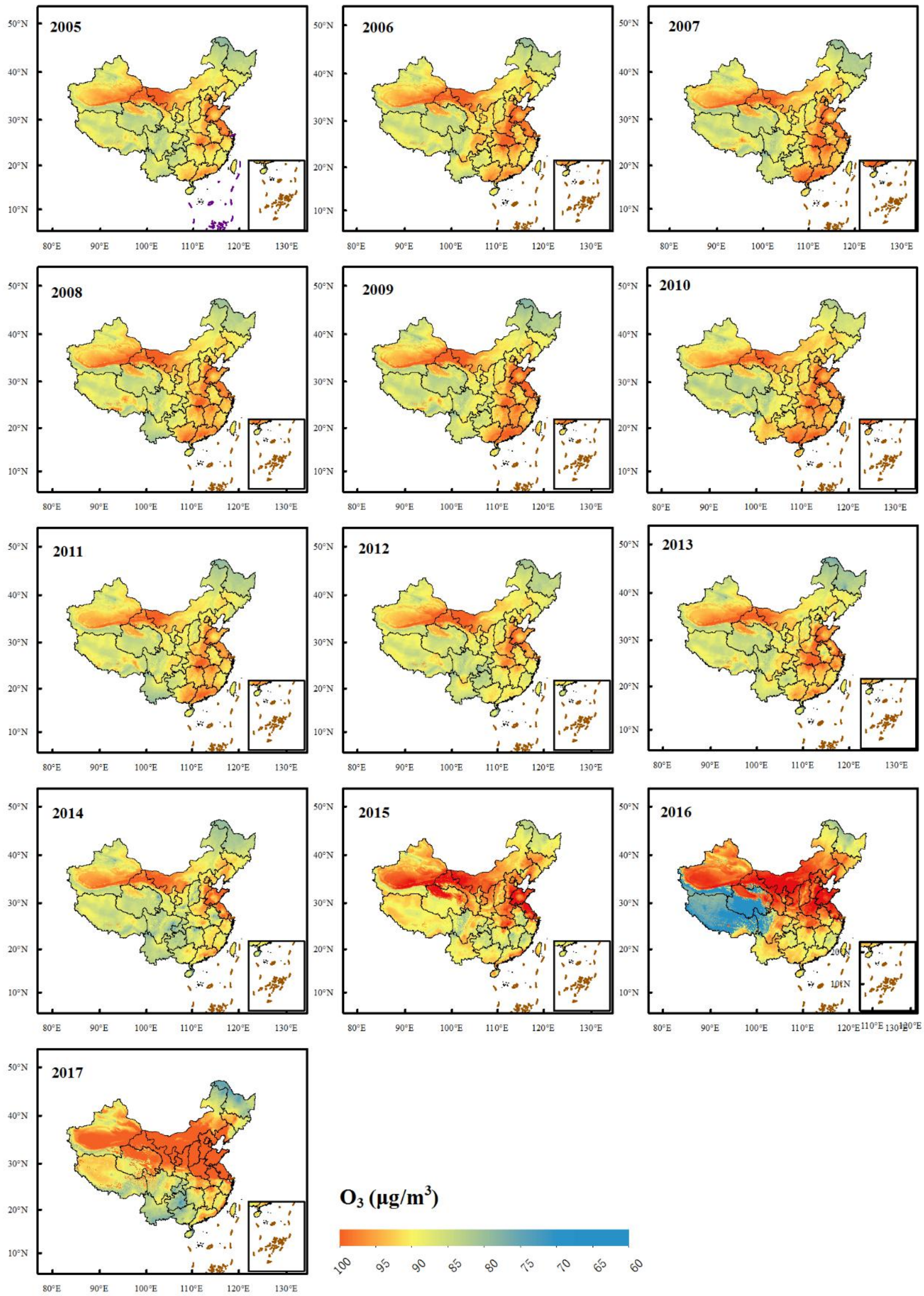

Fig. S2 Simulated annual mean O3-8hmax concentration based on random forest model in China from 2005 to 2017

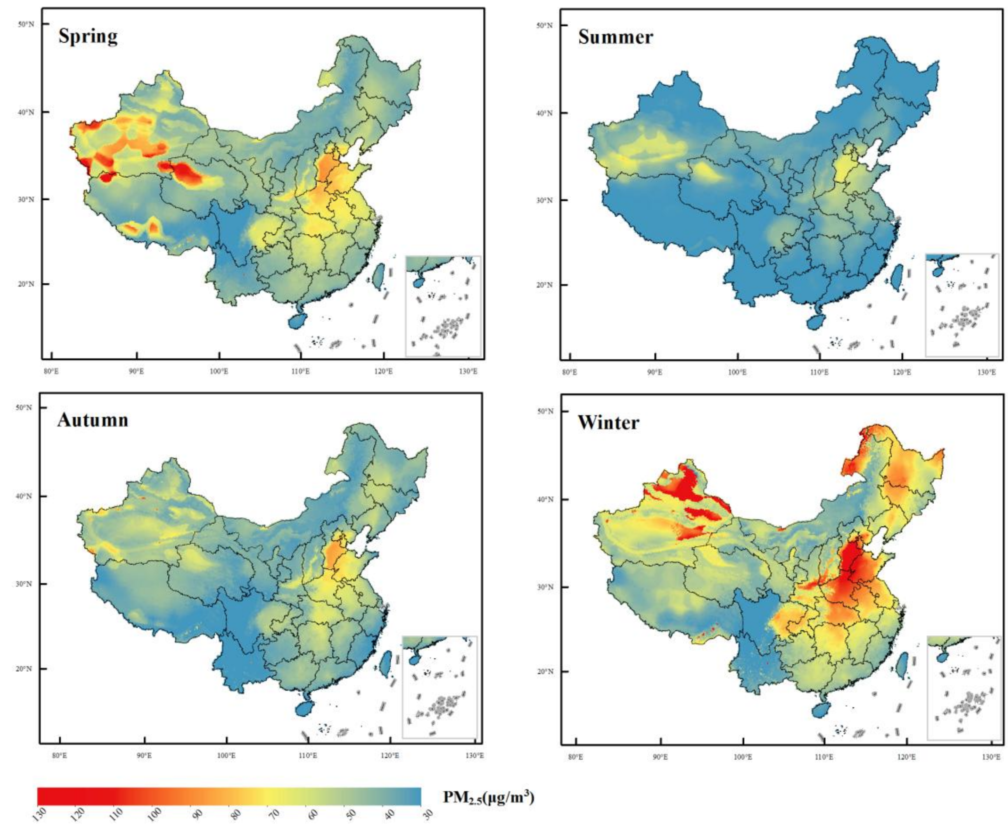

Fig. S3 Simulated seasonal mean PM2.5 concentration based on random forest model in China during 2005 to 2017

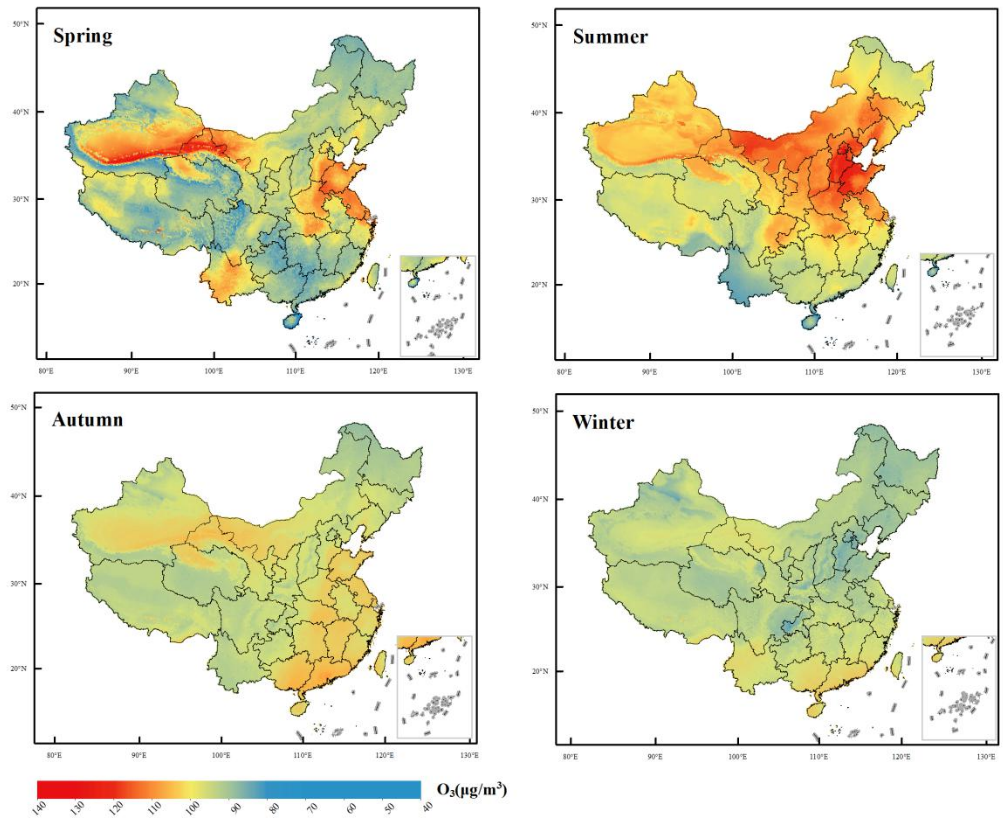

Fig. S4 Simulated seasonal mean O3-8hmax concentration based on random forest model in China during 2005 to 2017

References

Adam-Poupart, A. *et al.* (2014) 'Spatiotemporal Modeling of Ozone Levels in Quebec (Canada): A Comparison of Kriging, Land-Use Regression (LUR), and Combined Bayesian Maximum Entropy–LUR Approaches', *Environmental Health Perspectives*, 122(9), pp. 970–976. doi:10.1289/ehp.1306566.

Beelen, R. *et al.* (2009) 'Mapping of background air pollution at a fine spatial scale across the European Union', *Science of The Total Environment*, 407(6), pp. 1852–1867. doi:10.1016/j.scitotenv.2008.11.048.

Chen, G. *et al.* (2018) 'A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information', *Science of The Total Environment*, 636, pp. 52–60. doi:10.1016/j.scitotenv.2018.04.251.

Chen, Z.-Y. *et al.* (2019) 'Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China', *Atmospheric Environment*, 202, pp. 180–189. doi:10.1016/j.atmosenv.2019.01.027.

Di, Q. *et al.* (2017) 'A hybrid model for spatially and temporally resolved ozone exposures in the continental United States', *Journal of the Air & Waste Management Association*, 67(1), pp. 39–52. doi:10.1080/10962247.2016.1200159.

Fang, X. *et al.* (2016) 'Satellite-based ground PM2.5 estimation using timely structure adaptive modeling', *Remote Sensing of Environment*, 186, pp. 152–163. doi:10.1016/j.rse.2016.08.027.

Geng, G. *et al.* (2021) 'Tracking Air Pollution in China: Near Real-Time PM2.5 Retrievals from Multiple Data Sources', p. 28.

Hammer, M.S. *et al.* (2020) 'Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998–2018)', *Environmental Science & Technology*, 54(13), pp. 7879–7890. doi:10.1021/acs.est.0c01764.

Kerckhoffs, J. *et al.* (2015) 'A national fine spatial scale land-use regression model for ozone', *Environmental Research*, 140, pp. 440–448. doi:10.1016/j.envres.2015.04.014.

Li, T. *et al.* (2017) 'Estimating Ground-Level PM $_{2.5}$ by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach: Deep Learning for PM $_{2.5}$ Estimation', *Geophysical Research Letters*, 44(23), p. 11,985-11,993. doi:10.1002/2017GL075710.

Li, T. and Cheng, X. (2021) 'Estimating daily full-coverage surface ozone concentration using satellite observations and a spatiotemporally embedded deep learning approach', *International Journal of Applied Earth Observation and Geoinformation*, 101, p. 102356. doi:10.1016/j.jag.2021.102356.

Liu, R. *et al.* (2020) 'Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach', *Environment International*, 142, p. 105823. doi:10.1016/j.envint.2020.105823.

Ma, R. *et al.* (2021) 'Random Forest Model based Fine Scale Spatiotemporal O3 Trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017', *Environmental Pollution*, (10159), p. 116635.

Ma, Z. *et al.* (2014) 'Estimating Ground-Level PM $_{2.5}$ in China Using Satellite Remote Sensing', *Environmental Science & Technology*, 48(13), pp. 7436–7444. doi:10.1021/es5009399.

Ma, Z. *et al.* (2016) 'Satellite-Based Spatiotemporal Trends in PM $_{2.5}$ Concentrations: China, 2004–2013', *Environmental Health Perspectives*, 124(2), pp. 184–192. doi:10.1289/ehp.1409481.

Malmqvist, E. *et al.* (2014) 'Assessing ozone exposure for epidemiological studies in Malmö and Umeå, Sweden', *Atmospheric Environment*, 94, pp. 241–248. doi:10.1016/j.atmosenv.2014.05.038.

Son, Y. *et al.* (2018) 'Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters', *Science of The Total Environment*, 639, pp. 40–48. doi:10.1016/j.scitotenv.2018.05.144.

Wang, M. *et al.* (2015) 'Development of long-term spatiotemporal models for ambient ozone in six metropolitan regions of the United States: The MESA Air study', *Atmospheric Environment*, 123, pp. 79–87. doi:10.1016/j.atmosenv.2015.10.042.

Wang, M. *et al.* (2016) 'Combining Land-Use Regression and Chemical Transport Modeling in a Spatiotemporal Geostatistical Model for Ozone and PM $_{2.5}$', *Environmental Science & Technology*, 50(10), pp. 5111–5118. doi:10.1021/acs.est.5b06001.

Wei, J. *et al.* (2019) 'Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach', *Remote Sensing of Environment*, 231, p. 111221. doi:10.1016/j.rse.2019.111221.

Wei, J. *et al.* (2020) 'Improved 1 km resolution PM_{2.5} estimates across China using enhanced space–time extremely randomized trees', *Atmospheric Chemistry and Physics*, 20(6), pp. 3273–3289. doi:10.5194/acp-20-3273-2020.

Wei, J. *et al.* (2021) 'Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications', *Remote Sensing of Environment*, 252, p. 112136. doi:10.1016/j.rse.2020.112136.

Wolf, K. *et al.* (2017) 'Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany', *Science of The Total Environment*, 579, pp. 1531–1540. doi:10.1016/j.scitotenv.2016.11.160.

Xue, T. *et al.* (2019) ‘Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations’, *Environment International*, 123, pp. 345–357. doi:10.1016/j.envint.2018.11.075.

Yu, W. *et al.* (2017) ‘Improving satellite-based PM2.5 estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting’, *Scientific Reports*, 7(1), p. 7048. doi:10.1038/s41598-017-07478-0.

Zhan, Y. *et al.* (2018) ‘Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment’, *Environmental Pollution*, 233, pp. 464–473. doi:10.1016/j.envpol.2017.10.029.

Zhang, X.Y. *et al.* (2020) ‘Estimating Ground-Level Ozone Concentrations in Eastern China Using Satellite-Based Precursors’, *IEEE Transactions on Geoscience and Remote Sensing*, 58(7), pp. 4754–4763. doi:10.1109/TGRS.2020.2966780.