Supporting Information

Combining machine learning and numerical simulation for high-resolution $PM_{2.5}$ concentration forecast

Jianzhao Bi[1,**], K. Emma Knowland[2,3], Christoph A. Keller[2,3], Yang Liu[4,*]

[1]Department of Environmental & Occupational Health Sciences, University of Washington, Seattle, WA, USA
[2]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[3]Universities Space Research Association, Columbia, MD, USA
[4]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

Corresponding Authors
[*]Mailing Address: Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA. E-mail: yang.liu@emory.edu.
[**]Mailing Address: Department of Environmental & Occupational Health Sciences, University of Washington, 4225 Roosevelt Way NE, Seattle, WA 98105, USA. E-mail: jbi6@uw.edu.

Equations

Tables

Figures

Eq. S1:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Eq. S2:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Eq. S3:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Eq. S4:

$$NMB = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{\sum_{i=1}^{N} y_i}$$

Eq. S5:

$$PPV = \frac{\text{sensitivity}^{a} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

Eq. S6:

$$NPV = \frac{\text{specificity}^{b} \times (1 - \text{prevalence}^{c})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}$$

[a]Sensitivity: the proportion of days with a positive forecast result out of days with the actual pollution condition.
[b]Specificity: the proportion of days with a negative forecast result out of days without the pollution condition.
[c]Prevalence: the proportion of days with a particular pollution condition.

Table S1: Independent variables (predictors) used to build the RF-based forecast model (*s*: spatially varying at a 1-km resolution; *t*: temporally varying at the daily level).

$PM_{2.5}$-related variables
GEOS-CF $PM_{2.5}$ forecast concentration$_{(s,t)}$
$PM_{2.5}$ convolutional layer$_{(s,t)}$
Meteorological variables
Total cloud area fraction$_{(s,t)}$
Surface pressure$_{(s,t)}$
10-m specific humidity$_{(s,t)}$
10-m air temperature$_{(s,t)}$
Total precipitation$_{(s,t)}$
Tropopause pressure based on blended estimate$_{(s,t)}$
Surface skin temperature$_{(s,t)}$
10-m eastward wind$_{(s,t)}$
10-m northward wind$_{(s,t)}$
Planetary boundary layer height$_{(s,t)}$
Land-use variables
Population$_{(s)}$
Percentage of vegetation cover$_{(s)}$
Percentage of urban areas$_{(s)}$
Nearest distance to primary roads$_{(s)}$
Nearest distance to secondary roads$_{(s)}$

Table S2: The overall validation performance of our RF-based forecast models with different rolling periods (10, 30, and 60 days) for the validation period of March 11th, 2019 to March 10th, 2020.

Rolling Period	Forecast Day	N of Test Sample	R^2	RMSE ($\mu g/m^3$)	MAPE (%)	NMB
10 Days	Day 1	78378	0.71	20.42	37.7	0.012
	Day 2	78398	0.56	25.37	48.0	0.017
	Day 3	78158	0.48	27.76	52.3	0.020
	Day 4	78372	0.45	28.55	53.9	0.022
	Day 5	78372	0.43	29.26	56.6	0.027
30 Days	Day 1	78378	0.75	19.18	35.1	0.004
	Day 2	78398	0.62	23.56	44.2	0.009
	Day 3	78158	0.55	25.81	47.7	0.007
	Day 4	78372	0.51	26.89	49.8	0.002
	Day 5	78372	0.46	28.21	53.8	0.005
60 Days	Day 1	78378	0.76	18.70	34.3	0.003
	Day 2	78398	0.64	23.07	43.2	0.008
	Day 3	78158	0.56	25.48	46.8	0.005
	Day 4	78372	0.51	26.70	48.9	-0.001
	Day 5	78372	0.47	27.81	52.4	-0.003

Table S3: The overall validation performance of our RF-based forecast models with different rolling periods (60 and 90 days) for the validation period of April 2nd, 2019 to March 10th, 2020.

Rolling Period	Forecast Day	N of Test Sample	R^2	RMSE ($\mu g/m^3$)	MAPE (%)	NMB
60 Days	Day 1	73650	0.77	19.03	34.4	-0.001
	Day 2	73670	0.64	23.51	43.3	0.001
	Day 3	73430	0.56	25.99	46.9	-0.004
	Day 4	73644	0.52	27.15	48.7	-0.011
	Day 5	73644	0.48	28.22	52.0	-0.015
90 Days	Day 1	73650	0.77	18.91	34.4	0.002
	Day 2	73226	0.65	23.36	42.9	0.002
	Day 3	72543	0.57	25.92	46.4	-0.002
	Day 4	72366	0.52	27.17	48.4	-0.007
	Day 5	71936	0.50	28.02	51.3	-0.012

Table S4: The site-specific validation performance of our forecast framework (RF + GEOS-CF) and the original CTM forecast model (GEOS-CF) for the validation period of March 11th, 2019 to March 10th, 2020. The rolling period was 60 days.

Forecast Day	N of Sites	R^2	RMSE ($\mu g/m^3$)	MAPE (%)	NMB
RF + GEOS-CF					
Day 1	225	0.76 (0.08)*	18.25 (6.75)	33.1 (8.5)	0.009 (0.089)
Day 2	225	0.63 (0.09)	23.05 (8.31)	41.1 (11.1)	0.011 (0.115)
Day 3	225	0.54 (0.10)	25.60 (9.02)	44.1 (12.2)	0.005 (0.121)
Day 4	225	0.50 (0.08)	26.82 (10.53)	46.6 (13.2)	0.003 (0.134)
Day 5	225	0.45 (0.09)	28.12 (10.25)	49.4 (13.3)	-0.003 (0.144)
GEOS-CF					
Day 1	225	0.56 (0.21)	131.52 (78.03)	284.2 (102.3)	2.318 (0.971)
Day 2	225	0.55 (0.19)	125.10 (76.04)	280.7 (99.8)	2.274 (0.918)
Day 3	225	0.53 (0.18)	128.73 (74.02)	284.7 (92.1)	2.272 (0.859)
Day 4	225	0.50 (0.15)	124.90 (75.09)	286.4 (101.1)	2.251 (0.938)
Day 5	225	0.45 (0.16)	119.44 (71.46)	282.9 (98.3)	2.169 (0.962)

*Median (Interquartile Range/IQR)

Table S5: The day-specific validation performance of our forecast framework (RF + GEOS-CF) and the original CTM forecast model (GEOS-CF) for the validation period of March 11th, 2019 to March 10th, 2020. The rolling period was 60 days.

Forecast Day	N of Test Sample	R^2	RMSE (μg/m^3)	MAPE (%)	NMB
RF + GEOS-CF					
Day 1	357	0.53 (0.29)*	12.62 (11.28)	28.2 (16.5)	0.016 (0.236)
Day 2	357	0.38 (0.31)	15.29 (13.61)	34.5 (22.7)	0.031 (0.338)
Day 3	356	0.32 (0.33)	16.18 (14.82)	37.9 (25.0)	0.045 (0.379)
Day 4	357	0.30 (0.30)	16.95 (14.52)	39.8 (27.3)	0.035 (0.404)
Day 5	357	0.24 (0.30)	17.67 (15.51)	41.8 (29.3)	0.057 (0.460)
GEOS-CF					
Day 1	357	0.33 (0.34)	101.43 (81.28)	252.3 (136.8)	2.255 (1.307)
Day 2	357	0.30 (0.31)	100.64 (81.11)	250.9 (142.8)	2.242 (1.195)
Day 3	356	0.27 (0.33)	98.63 (79.59)	258.8 (148.2)	2.315 (1.265)
Day 4	357	0.23 (0.33)	99.48 (75.19)	256.5 (153.1)	2.321 (1.271)
Day 5	357	0.20 (0.29)	95.38 (79.65)	253.6 (166.0)	2.176 (1.413)

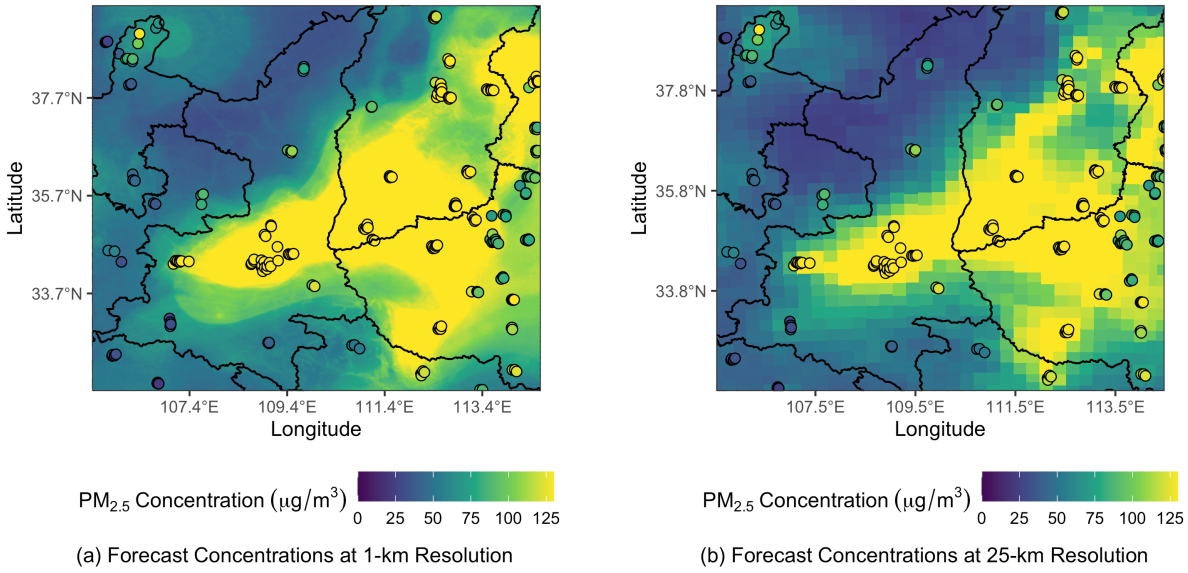
*Median (Interquartile Range/IQR)

Table S6: The overall validation performance of our forecast framework (RF + GEOS-CF) with 25-km predictors for the validation period of March 11th, 2019 to March 10th, 2020. The rolling period was 60 days.

Forecast Day	N of Test Sample	R^2	RMSE ($\mu g/m^3$)	MAPE (%)	NMB
Day 1	78378	0.75	19.01	35.4	0.012
Day 2	78398	0.63	23.19	43.9	0.021
Day 3	78158	0.56	25.39	47.2	0.016
Day 4	78372	0.52	26.63	49.2	0.011
Day 5	78372	0.47	27.91	52.4	0.010

Table S7: The overall validation performance of the RF-based versus XGBoost-based forecast models for November 2019. The rolling period was 60 days.

Forecast Day	N of Test Sample	RF		XGBoost	
		R^2	MAPE (%)	R^2	MAPE (%)
Day 1	6628	0.70	35.0	0.66	36.8
Day 2		0.55	39.4	0.55	40.0
Day 3		0.52	38.6	0.51	38.9
Day 4		0.46	39.1	0.46	40.4
Day 5		0.38	42.4	0.43	43.2



(a) Forecast Concentrations at 1-km Resolution

(b) Forecast Concentrations at 25-km Resolution

Figure S1: Spatial $PM_{2.5}$ forecast concentrations at (a) 1-km and (b) 25-km resolutions based on the models trained with predictors at the corresponding resolutions. These were the next-day concentrations forecasted on January 24th, 2020 (*i.e.*, for January 25th, 2020). The colored dots on the 1-km subfigure show the observed $PM_{2.5}$ concentrations at the monitoring locations, which share the same color scheme with the forecast concentrations.