Supplementary information:

Cooperative simultaneous inversion of satellite-based real-time $PM_{2.5}$ and ozone levels using an improved deep learning model with attention mechanism

Xing Yan, Chen Zuo, Zhanqing Li, Hans W. Chen*,Yize Jiang, Bin He, Huiming Liu, Jiayi Chen, Wenzhong Shi

Table S1. Summary of the data sources used in this study.

Variable	Description	Unit	Spatial Resolution	Temporal Resolution	Data Source
MOD02SSH	Radiance	$W\ m^{-2}\ \mu m^{-1}\ sr^{-1}$	5 km	Daily	LAADS DAAC (https://ladsweb.modaps.eosdis.nasa.gov/)
MOD09	Surface reflectance	-	0.05°×0.05°	Daily	
MOD13	Normalized difference vegetation index	-	0.05°×0.05°	16-Day	
MOD11	Land surface temperature	K	1000 m	Daily	
MCD12	Landcover	Land cover type	500 m	Annual	
GEOS-CF	On-going 5-day forecasts of $PM_{2.5}$	$\mu m\ m^{-3}$	0.3°×0.25°	3-Hour	GMAO (https://gmao.gsfc.nasa.gov/)
	On-going 5-day forecasts of Ozone	$\mu m\ m^{-3}$			
RH	Relative humidity	%		Hourly	National Climatic Data Center (https://www.ncdc.noaa.gov/)
T2M	2-m air temperature	K			
WS	Wind speed	$m\ s^{-1}$			
BLH	Boundary layer height	m			
Visibility	Visibility data observed from meteorological stations	m			
$PM_{2.5}$	Ground-based $PM_{2.5}$	$\mu m\ m^{-3}$		Hourly	CNEMC (http://www.cnemc.cn)
O_3	Ground-based O3	$\mu m\ m^{-3}$			

Table S2. Summary of the input Variables in SOPiNet.

Variable	Class	Name	Detail Description
Continuous variables	Satellite Data	B1-B12, B17-B36	Band 1-36, solar and satellite zenith angle, and azimuth angles
		B1-B7	Surface reflectance
		NDVI	Normalized difference vegetation index
		LST	Land surface temperature
	Global Modeling and Assimilation	GEOS-PM2.5	Real-time $PM_{2.5}$ concentration
		GEOS-O3	Real-time O_3 concentration
	Ground-based measurement data	Past-PM2.5	$PM_{2.5}$ in the past 20 days
		Past-O3	O_3 in the past 20 days
		BLH	Boundary layer height
		T2M	2-m air temperature
		WS	Wind speed
		RH	Relative humidity
Categorical variables		LC	Land cover
		Prov	Province region
		Month	
		Season	
		P-M	Province-month pairwise feature

Table S3. The class types and their abbreviations of land use type in Figure S2.

Value	Land use type
1	Evergreen Needleleaf Forests
2	Evergreen Broadleaf Forests
3	Deciduous Needleleaf Forests
4	Deciduous Broadleaf Forests
5	Mixed Forests
6	Closed Shrublands
7	Open Shrublands
8	Woody Savannas
9	Savannas
10	Grasslands
11	Permanent Wetlands
12	Croplands
13	Urban and Built-up Lands
14	Cropland/Natural Vegetation Mosaics
15	Permanent Snow and Ice
16	Barren
17	Water Bodies

Table S4. The class types and their abbreviations of province in Figure S2.

Value	Province	Value	Province
1	Zhejiang	18	Hubei
2	Yunnan	19	Heilongjiang
3	Xinjiang	20	Henan
4	Xianggang	21	Beijing
5	Xizang	22	Tianjin
6	Taiwan	23	Hainan
7	Sichuan	24	Guizhou
8	Shaanxi	25	Guangxi
9	Shanxi	26	Gansu
10	Shandong	27	Fujian
11	Qinghai	28	Aomen
12	Ningxia	29	Anhui
13	NeiMongol	30	Shanghai
14	Liaoning	31	Chongqing
15	Jiangxi	32	Jiangsu
16	Jilin	33	Guangdong
17	Hunan	34	Hebei

Table S5. Previous studies in China. We only use the time-based validation method result for comparison.

	Cooperative inversion	Full coverage	Model	Spatial resolution	Validation method	R^2	RMSE(μg/m^3)	Literature
$PM_{2.5}$	no	no	STET	1 km	train:2018 test:2017	0.65	-	Wei et al.(2020)[1]
	no	no	JFRF	5 km	train:2018 test:2019	0.70	18.39	Dong et al.(2022)[2]
	no	no	EntityDenseNet	5 km	train:2016-2018 test:2019	0.65	25.30	Yan et al.(2020)[3]
	no	no	SIDLM	3 km	train:2016-2018 test:2019	0.70	15.30	Yan et al.(2021)[4]
	no	yes	TAP $PM_{2.5}$	10 km	Time-based CV	0.58	27.50	Geng et al.(2022)[5]
	no	yes	NSTC model	1 km	Time-based CV	0.62	27.70	Huang et al.(2022)[6]
	no	yes	ML-CMAQ	10 km	train:2013-2016 test:2017-2019	0.62	27.80	Xue et al.(2019)[7]
	yes	**yes**	**SOPiNet**	**5 km**	**train:2019-2021 test:2022**	**0.72**	**16.45**	**This study**
O_3	no	no	Semi-SIDLM	5 km	train:2016-2018 test:2019	0.71	21.88	Yan et al.(2021)[8]
	no	no	ExDLM	5km	Time-based CV	0.78	18.35	Luo et al.(2022)[9]
	no	no	XGBoost	10 km	Time based CV	0.78	21.47	Liu et al(2020)[10]
	no	no	SGLboost	2km	train:2017-2018 test:2019	0.72	25.11	Wang et al.(2022)[11]
	yes	**yes**	**SOPiNet**	**5km**	**train:2019-2021 test:2022**	**0.82**	**12.60**	**This study**

Table S6. The training and inference time

Model	Training (each epoch)	Inference time
SOPiNet ($PM_{2.5}$+O_3)	89±3 s	17±2 s
Single Modeling ($PM_{2.5}$) + Single Modeling (O_3)	138±7 s	25±2 s

per epoch (mean ± std. dev. of 100 epoch)

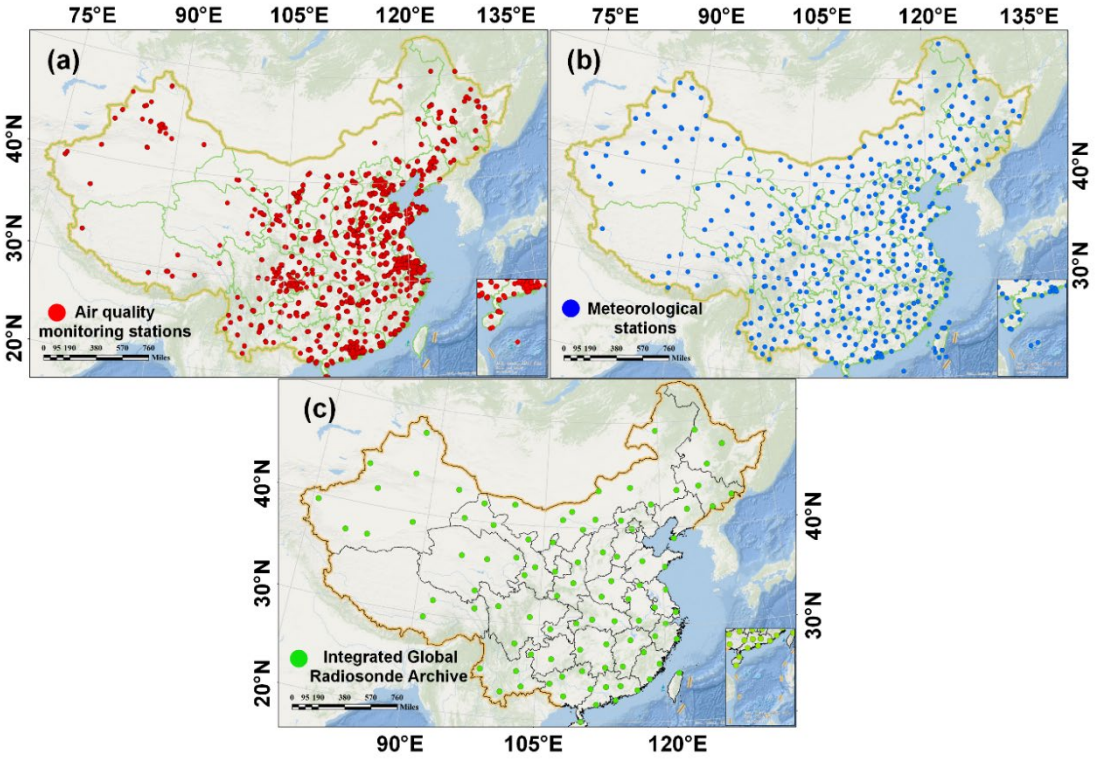

Figure S1. (a) CNEMC air quality monitors in China. (b) Spatial distribution of the meteorological stations. (c) The Integrated Global Radiosonde Archive sites in China.

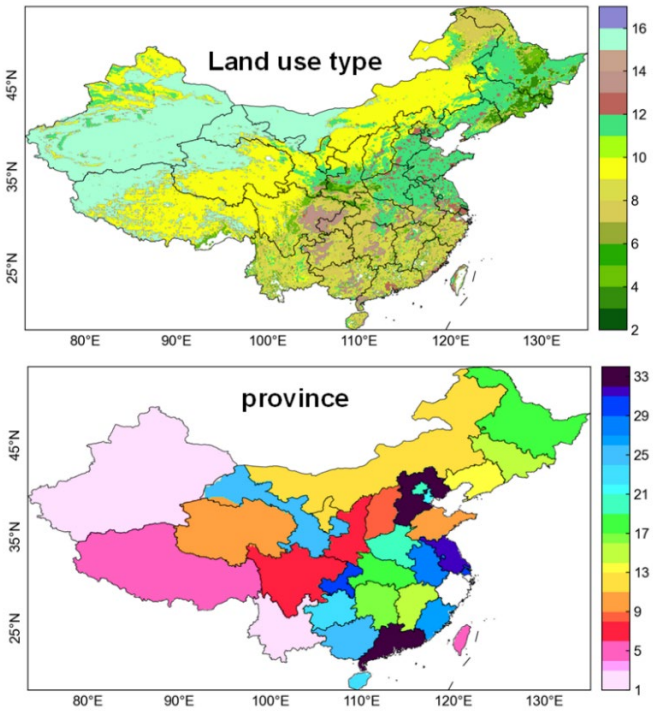

Figure S2. Spatial distribution of the geographical variables used in this study. The detailed descriptions for their legend are in Tables S2-S3.

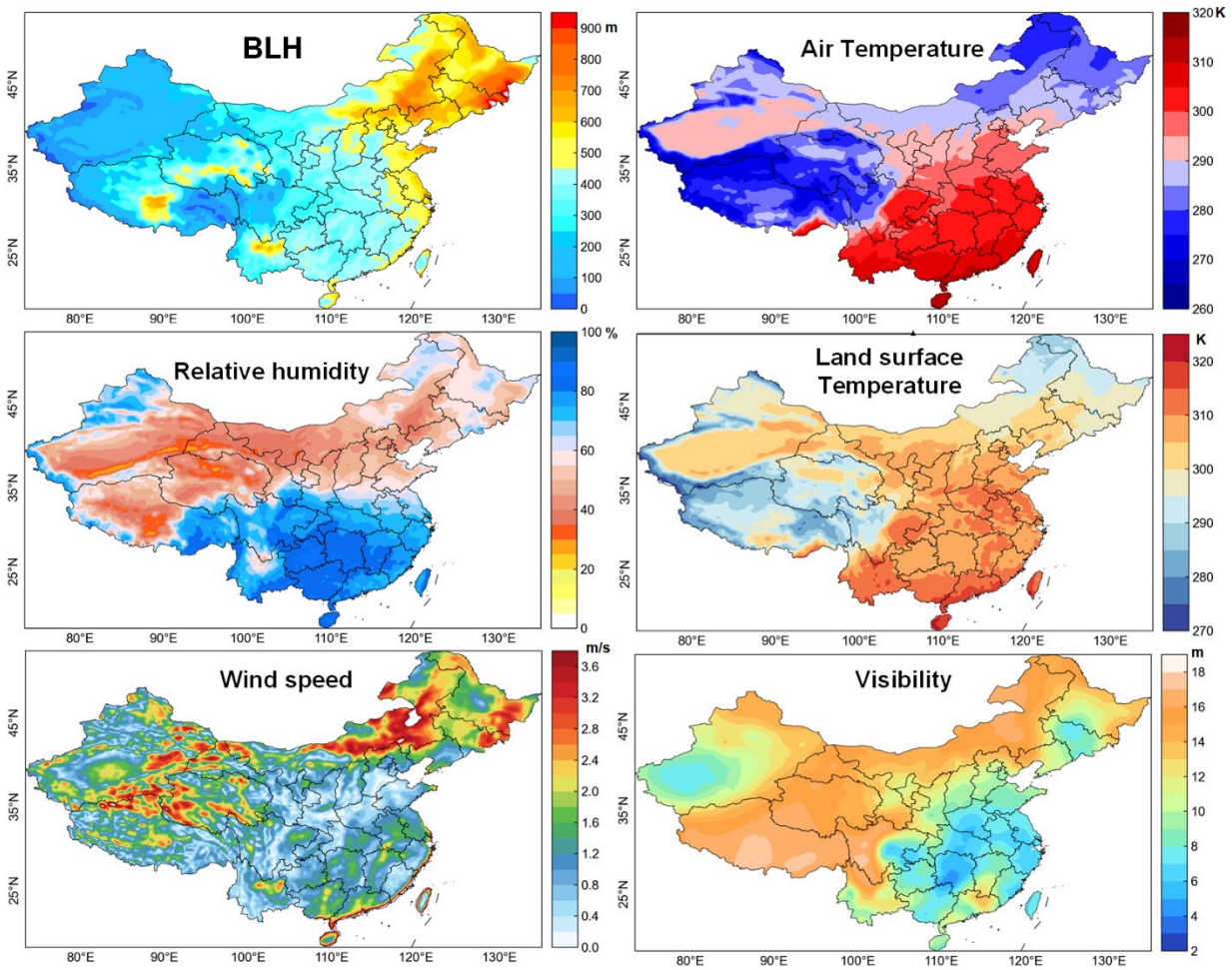

Figure S3. Spatial distribution of the meteorological variables during 2019-2022.

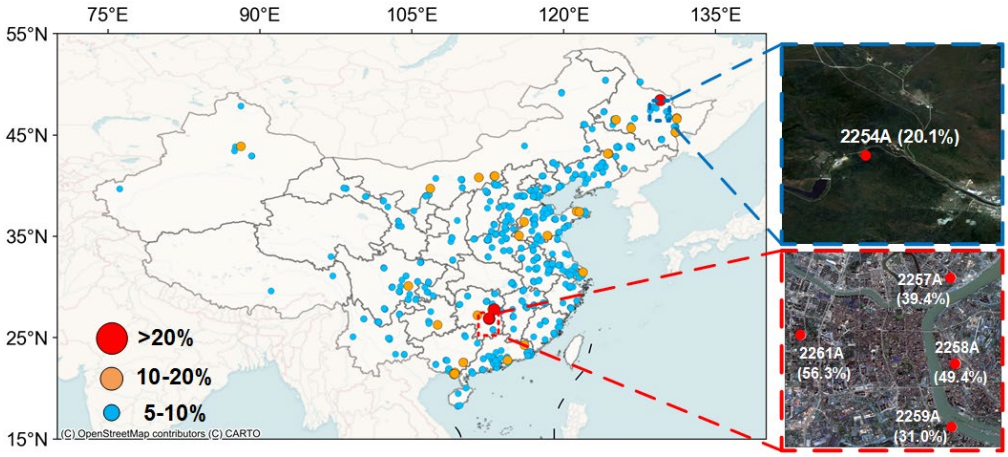

Figure S4. Spatial distribution of the sites with missing data in three intervals. The detailed descriptions for their information can be found in Table S4.

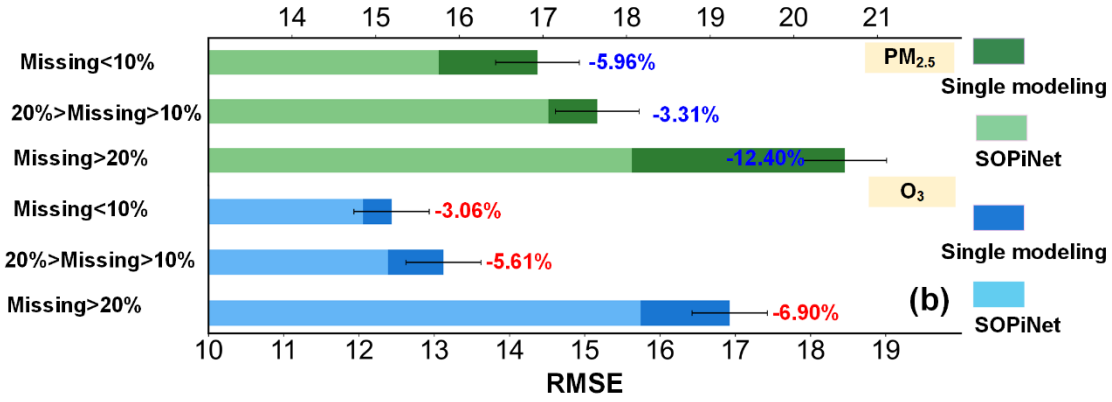

Figure S5. The RMSE validation of SOPiNet and single modeling for different levels of missing data.

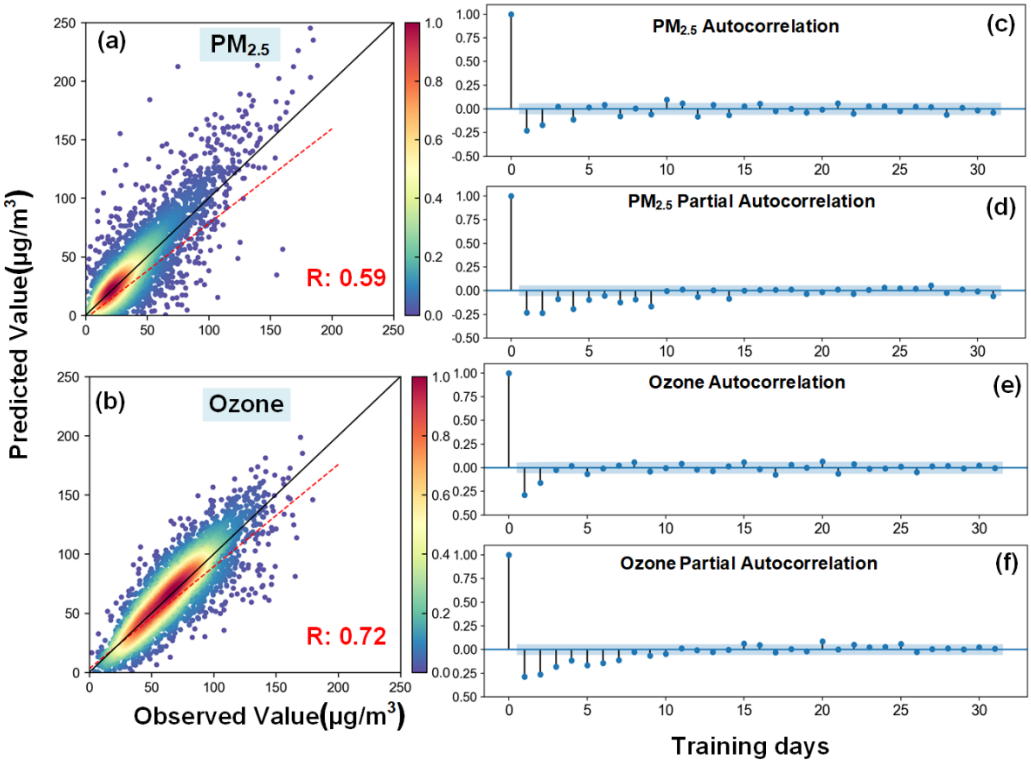

Figure S6. Evaluation of ARIMA model of $PM_{2.5}$ (a) and O_3 (b). (c-f) ACF and PACF plots to determine the p and q parameters in the ARIMA model.

Optimization of the number of past days' information for $PM_{2.5}$ and O_3

Information about O_3 and $PM_{2.5}$ from previous days is crucial for the estimation of real-time O_3 and $PM_{2.5}$ by the developed SOPiNet. As seen in Figures S7 a and b, the linear correlation coefficients between current $PM_{2.5}$ or O_3 concentrations with their past values range from 0.66 to 0.39 and 0.78 to 0.55, respectively. However, the correlation coefficient cannot directly aid in the inference of the optimal number of past time periods to include for $PM_{2.5}$ and O_3 due to autocorrelations between nearby days. Here, we used the Autoregressive Integrated Moving Average (ARIMA) model to

determine the number of past days to use as input to SOPiNet. The model is expressed as ARIMA(p, d, q), where the parameters p, d, and q determine the structure of the model, which is a combination of auto-regression AR(p), moving average MA(q), and differencing degree d. The mathematical formula for the ARIMA(p, d, q) model can be expressed as follows:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \quad (12)$$

where L is the lag operator, ϕ_i is the parameter for the auto-regressive part of the model, θ_i is the parameter for the moving average part, and ε_i denotes error terms.

To determine the optimum number of input days, we fitted ARIMA models to the $PM_{2.5}$ and O_3 concentrations in 2019–2020 using a varying number of past days in the models, and validated the results during 2021. We used a first degree of differencing to construct stationary time series and determined the optimal parameter values for p and q based on the Akaike Information Criterion using the Auto-ARIMA function (the detailed can be found in Supplementary information-Determine the parameter for ARIMA model, Figure S8). Figures S7c and d present the correlation coefficient (R) and root-mean-square error (RMSE) average values of each ARIMA-based $PM_{2.5}$ and O_3 modeling results for different numbers of past days as input. The red dots in the graph indicate the R of the modeling results relative to the number of past days and the box line diagram shows the corresponding RMSE. For O_3, the results show a clear peak in R when using the past 20 days as input, with R being 0.73 and RSME being 12.19 $\mu g/m^3$. As for $PM_{2.5}$, when including information on the past 20 days, R is 0.60 and RMSE is 21.82 $\mu g/m^3$. Therefore, based on these results, we chose to use 20 days of

information as inputs for the SOPiNet real-time O_3 and $PM_{2.5}$ retrievals.

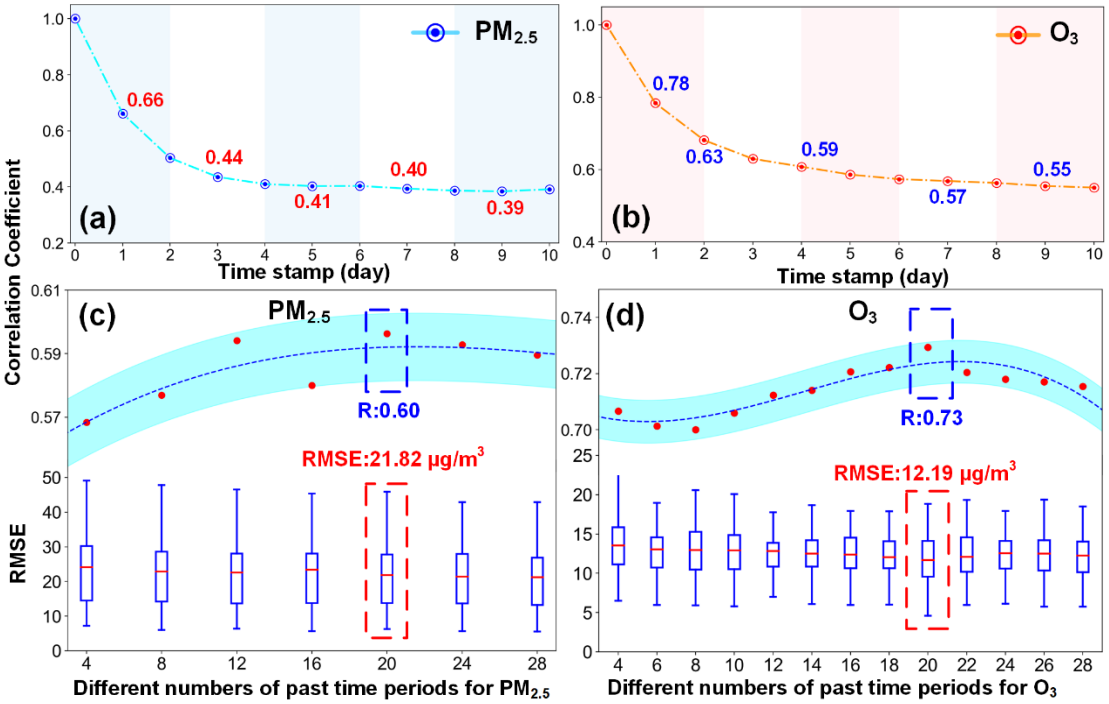

Figure S7. (a-b) Pearson correlation coefficient (R) between $PM_{2.5}$ (a) and O_3 (b) at the current time with values from the past 10 days. (c-d) Validation of predicted $PM_{2.5}$ and O_3 from the ARIMA model using different numbers of past days as training data. The red dots above show the correlations between the predicted and the measured $PM_{2.5}$ and O_3 concentrations. The blue dashed line shows the fitted results and the shading indicates half of the standard deviation. The box plots below show the root-mean-square error (RMSE) values for the predicted results. The dashed boxes highlight the results when using 20 days of training data, with the corresponding R (RMSE) values shown below (above) the boxes.

Determine the parameter for ARIMA model

To find the optimum number of input days, we tested a range of ARIMA models fitted using a varying number of past days from 2 to 30 days in intervals of 2 days. We constructed a sliding scale prediction based on the common ARIMA model. For example, when input steps is 30, we entered the ground observations from the last thirty days as training data and slide the prediction until all the data in the test datasets are defined (see Figure S7). For each ARIMA model, we automatically selected the optimal parameters using the auto-ARIMA package (https://github.com/alkaline-ml/pmdarima). There are three steps to determine the parameters[12-13]. To make the datasets stationary, the first step is differencing the time series. In this package, the number of differences d is determined using repeated KPSS tests. The values of p and q were then chosen by minimizing the AIC after differencing the data d times by a stepwise search to traverse the model space. Finally, the model with the smallest AIC value is chosen. To ensure that our data were usable with the ARIMA model, we first performed a manual test and determined the parameters using the partial autocorrelation function (PACF) for p and autocorrelation function (ACF) for q. The manual tests were consistent with the auto-ARIMA results.

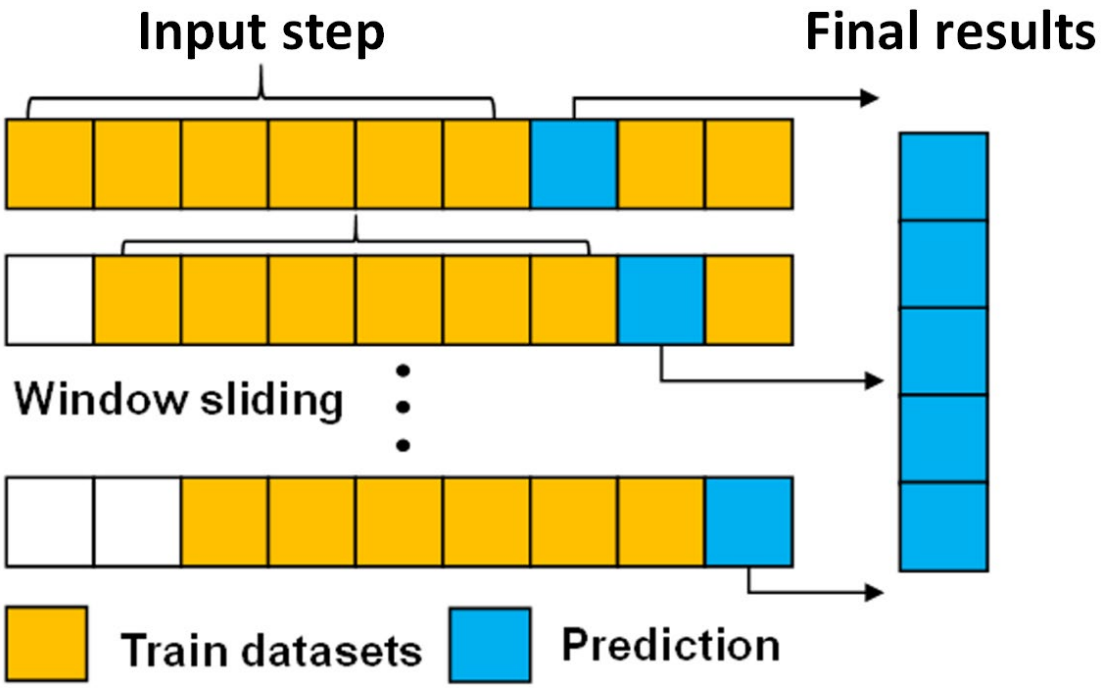

Figure S8. ARIMA flow chart of how to find the optimum number of input days.

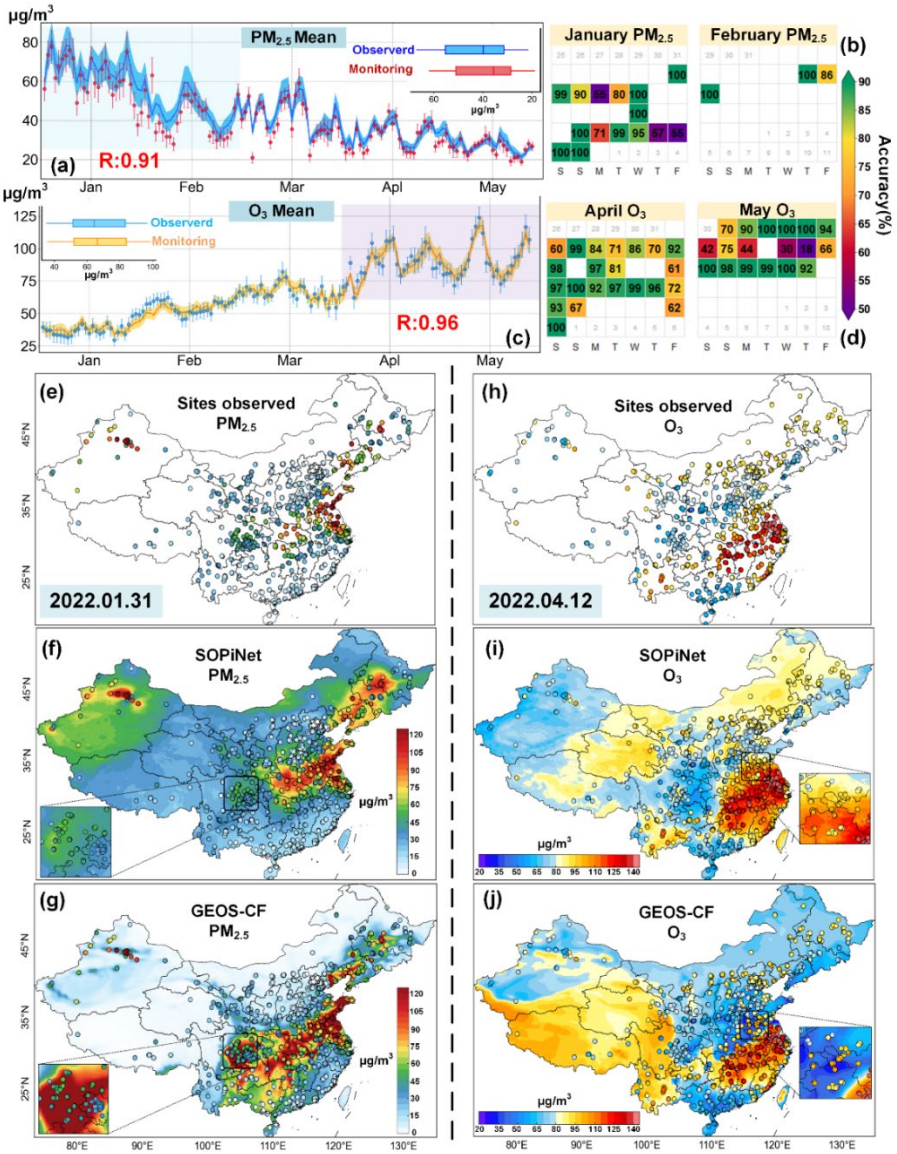

Figure S9. (a, c) Performance of SOPiNet real-time monitoring results in 2022. The lines represent the SOPiNet estimated daily mean, and the shadings show the plus/minus one standard deviation. Whiskers on dots are the daily observed value plus/minus one standard deviation. (b, d) Accuracy of SOPiNet alerts of $PM_{2.5}$ and O_3 in heavily polluted conditions (>100 μg/m^3), respectively. (e-j) Comparison of SOPiNet with GEOS-CF results in two cases. (e) and (h) show the in situ concentration measurements; (f) and (i) the SOPiNet retrievals; and (g) and (j) the GEOS-CF forecast values.

References:

1.J. Wei, Z. Q. Li, M. Cribb, W. Huang, W. H. Xue, L. Sun, J. P. Guo, Y. R. Peng, J. Li, A. Lyapustin, L. Liu, H. Wu andY. M. Song, *ATMOS CHEM PHYS*, 2020, **20**, 3273-3289.

2.L. C. Dong, S. W. Li, J. Xing, H. Lin, S. S. Wang, X. Y. Zeng andY. M. Qin, *ATMOS ENVIRON*, 2022, **273**.

3.X. Yan, Z. Zang, N. N. Luo, Y. Z. Jiang andZ. Q. Li, *ENVIRON INT*, 2020, **144**.

4.X. Yan, Z. Zang, Y. Z. Jiang, W. Z. Shi, Y. S. Guo, D. Li, C. F. Zhao andL. T. Husi, *ENVIRON POLLUT*, 2021, **273**.

5.G. N. Geng, Q. Y. Xiao, S. G. Liu, X. D. Liu, J. Cheng, Y. X. Zheng, T. Xue, D. Tong, B. Zheng, Y. R. Peng, X. M. Huang, K. B. He andQ. Zhang, *ENVIRON SCI TECHNOL*, 2021, **55**, 12106-12115.

6.C. H. Huang, J. L. Hu, T. Xue, H. Xu andM. Wang, *ENVIRON SCI TECHNOL*, 2021, **55**, 2152-2162.

7.T. Xue, Y. X. Zheng, D. Tong, B. Zheng, X. Li, T. Zhu andQ. Zhang, *ENVIRON INT*, 2019, **123**, 345-357.

8.Z. Zang, Y. S. Guo, Y. Z. Jiang, C. Zuo, D. Li, W. Z. Shi andX. Yan, *INT J APPL EARTH OBS*, 2021, **103**.

9.N. Luo, Z. Zang, C. Yin, M. Liu, Y. Jiang, C. Zuo, W. Zhao, W. Shi andX. Yan, *ATMOS ENVIRON*, 2022, **290**, 119370.

10.R. Y. Liu, Z. W. Ma, Y. Liu, Y. C. Shao, W. Zhao andJ. Bi, *ENVIRON INT*, 2020, **142**.

11.Y. Wang, Q. Q. Yuan, L. Y. Zhu andL. P. Zhang, *GEOSCI FRONT*, 2022, **13**.

12.D. Y. Fan, H. Sun, J. Yao, K. Zhang, X. Yan andZ. X. Sun, *ENERGY*, 2021, **220**.

13.H. F. Zou, F. F. Yang andG. P. Xia, *Time series forecasting model using a hybrid ARIMA and neural network*, 2005.