

## ARTICLE OPEN



# A synchronized estimation of hourly surface concentrations of six criteria air pollutants with GEMS data

Qianqian Yang<sup>1,2</sup>, Jhoon Kim<sup>3</sup>, Yeseul Cho<sup>3</sup>, Won-Jin Lee<sup>4</sup>, Dong-Won Lee<sup>4</sup>, Qiangqiang Yuan<sup>2</sup>, Fan Wang<sup>1</sup>, Chenhong Zhou<sup>5</sup>, Xiaorui Zhang<sup>1</sup>, Xiang Xiao<sup>1</sup>, Meiyu Guo<sup>1</sup>, Yike Guo<sup>1,5</sup>, Gregory R. Carmichael<sup>6</sup> and Meng Gao<sup>1</sup>✉

Machine learning is widely used to infer ground-level concentrations of air pollutants from satellite observations. However, a single pollutant is commonly targeted in previous explorations, which would lead to duplication of efforts and ignorance of interactions considering the interactive nature of air pollutants and their common influencing factors. We aim to build a unified model to offer a synchronized estimation of ground-level air pollution levels. We constructed a multi-output random forest (MORF) model and achieved simultaneous estimation of hourly concentrations of PM<sub>2.5</sub>, PM<sub>10</sub>, O<sub>3</sub>, NO<sub>2</sub>, CO, and SO<sub>2</sub> in China, benefiting from the world's first geostationary air-quality monitoring instrument Geostationary Environment Monitoring Spectrometer. MORF yielded a high accuracy with cross-validated R<sup>2</sup> reaching 0.94. Meanwhile, model efficiency was significantly improved compared to single-output models. Based on retrieved results, the spatial distributions, seasonality, and diurnal variations of six air pollutants were analyzed and two typical pollution events were tracked.

*npj Climate and Atmospheric Science* (2023)6:94; <https://doi.org/10.1038/s41612-023-00407-1>

## INTRODUCTION

With rapid and energy-intensive economic development, China has witnessed serious air pollution in the past several decades<sup>1</sup>. To assist air quality management, the China National Environmental Monitoring Center (CNEMC) started in 2013 to operate a network that measures six criteria air pollutants, namely PM<sub>2.5</sub>, PM<sub>10</sub>, sulfur dioxide (SO<sub>2</sub>), nitrogen dioxide (NO<sub>2</sub>), carbon monoxide (CO), and ozone (O<sub>3</sub>). However, these sites are predominantly concentrated in urban or suburban regions, and considerable areas of China go still unmonitored<sup>2,3</sup>. Satellite-based observations, especially geostationary satellites, are skilled in offering horizontal distribution of atmospheric composition and are thus widely used to supplement ground-based observations<sup>4</sup>.

Many algorithms have been developed or adopted to retrieve ground-level abundance of air pollutants from satellite images, which can be roughly divided into physics-based and statistics-based methods. Physics-based approach converts column density measured from satellites to ground-level concentrations using their physical connections. For instance, ref. <sup>5</sup> used SO<sub>2</sub> profiles from an air quality model and tropospheric column SO<sub>2</sub> from the Ozone Monitoring Instrument (OMI) to estimate ground-level SO<sub>2</sub> concentrations. Similarly, ground-level NO<sub>2</sub> concentrations were inferred from NO<sub>2</sub> vertical column abundances from the TROPospheric Monitoring Instrument (TROPOMI) using a surface-to-column conversion factor from a chemical transport model<sup>6</sup>. Additionally, a semi-empirical physical approach was developed to obtain ground-level PM<sub>2.5</sub> and PM<sub>10</sub> from satellite aerosol optical depth (AOD) through vertical correction, humidity correction, fine mode conversion, and volume correction<sup>7</sup>.

The statistics-based approach aims at learning relationships between satellite retrievals of aerosols and gases and collocated ground-level concentration with statistical models. Due to skills in capturing nonlinear relationships, machine learning-based retrieval

has received considerable attention in recent years<sup>8–13</sup>. Wang and Christopher<sup>14</sup> found a linkage between AOD and ground-level PM<sub>2.5</sub> mass and estimated air quality categories from AOD using a linear regression model. Later, more influencing factors such as meteorological and topographical variables were considered and more advanced and sophisticated machine learning models, such as land-use regression model<sup>15</sup>, space-time regression model<sup>16</sup>, geointelligent deep neural networks<sup>17,18</sup>, and ensemble-learning-based models<sup>19,20</sup> were developed. For gaseous pollutants, satellite retrievals of column density were usually used as main predictors<sup>21–23</sup>.

Recently, retrieval models have also been built using satellite radiance or reflectance data. Shen et al.<sup>24</sup> proposed that we could replace AOD with top-of-atmosphere reflectance (TOAR) data for PM<sub>2.5</sub> estimation, and successfully retrieved ground-level PM<sub>2.5</sub> concentration with TOAR data from three MODIS bands (red, blue, and a short-wave infrared band). After that, studies based on Himawari-8, Fengyun-4, and Landsat-8 data<sup>25</sup> were conducted for ground-level PM<sub>2.5</sub> and PM<sub>10</sub> concentrations estimation. In addition, reflectance/radiance-based retrieval of ground-level O<sub>3</sub> concentration was also achieved. Luo et al.<sup>26</sup> utilized 32 MODIS wavebands (all MODIS bands excluded bands 13–16, from visible to thermal infrared bands) and other auxiliary variables to estimate ground-level O<sub>3</sub> concentration with a deep-learning technique. Similarly, ref. <sup>27</sup> leveraged 7 Himawari-8 channels (one mid-wave infrared and six thermal infrared channels) for O<sub>3</sub> estimation and also achieved good results. It was demonstrated that radiance/reflectance-based models achieved similar accuracy to column-product-based models but improved resolution and spatial coverage<sup>24,28,29</sup>.

A single specific pollutant was targeted in above mentioned models (referred to as single-output models hereafter). Single-output models are skilled in exploring the characteristics of a single pollutant fully, and high model accuracy was usually yielded. However, major air pollutants share some common

<sup>1</sup>Department of Geography, Faculty of Social Sciences, Hong Kong Baptist University, Hong Kong, SAR 999077, China. <sup>2</sup>School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei 430079, China. <sup>3</sup>Department of Atmospheric Sciences, Yonsei University, Seoul 03722, Korea. <sup>4</sup>Environmental Satellite Center, National Institute of Environmental Research, Incheon 22689, Korea. <sup>5</sup>Department of Computer Science, Faculty of Science, Hong Kong Baptist University, Hong Kong, SAR 999077, China. <sup>6</sup>Department of Chemical and Biochemical Engineering, The University of Iowa, Iowa City, IA 52242, USA. ✉email: mmgao2@hkbu.edu.hk

sources, evolve under the same meteorological conditions, and are connected chemically or physically<sup>30</sup>. Therefore, similar predictors and model structures were used in models built for different pollutants. Building multiple single-output models for different pollutants leads to duplication of efforts. A model that estimates concentrations of these pollutants simultaneously can make better use of their correlations and improve efficiency. Multi-output regression methods<sup>31</sup> that consider both underlying relationships between features and corresponding target variables and relationships between targets have been applied in ecological modeling<sup>32,33</sup>, chemometrics<sup>34</sup>, signal and image processing<sup>35,36</sup>, etc., and demonstrated a strong ability on simultaneous prediction and joint estimation of multiple variables.

Another limitation of previous studies is embedded in the usage of low-Earth orbiting satellite that provides one to two observations for concerned areas and misses the dynamic evolution of pollutants during a day<sup>37,38</sup>. Geostationary observations of AOD were made available since the launch of a geostationary meteorological satellite, yet those of trace gases have been limited as satellite monitoring of trace gases relies largely on spectral information at ultraviolet (UV) and visible bands. In February 2020, South Korea launched the Geostationary Environment Monitoring Spectrometer (GEMS) on board the Geostationary Korea Multi-Purpose Satellite 2 (GEO-KOMPSAT-2) satellite series. GEMS is the first ultraviolet-visible instrument onboard a geostationary earth orbit platform<sup>38</sup>, which enables hourly monitoring of trace gases for almost 20 countries in Asia. With the high spatial and temporal resolution, diurnal variations of multiple atmospheric components are observed, which provides a great opportunity to research hourly estimations of ground-level air pollution.

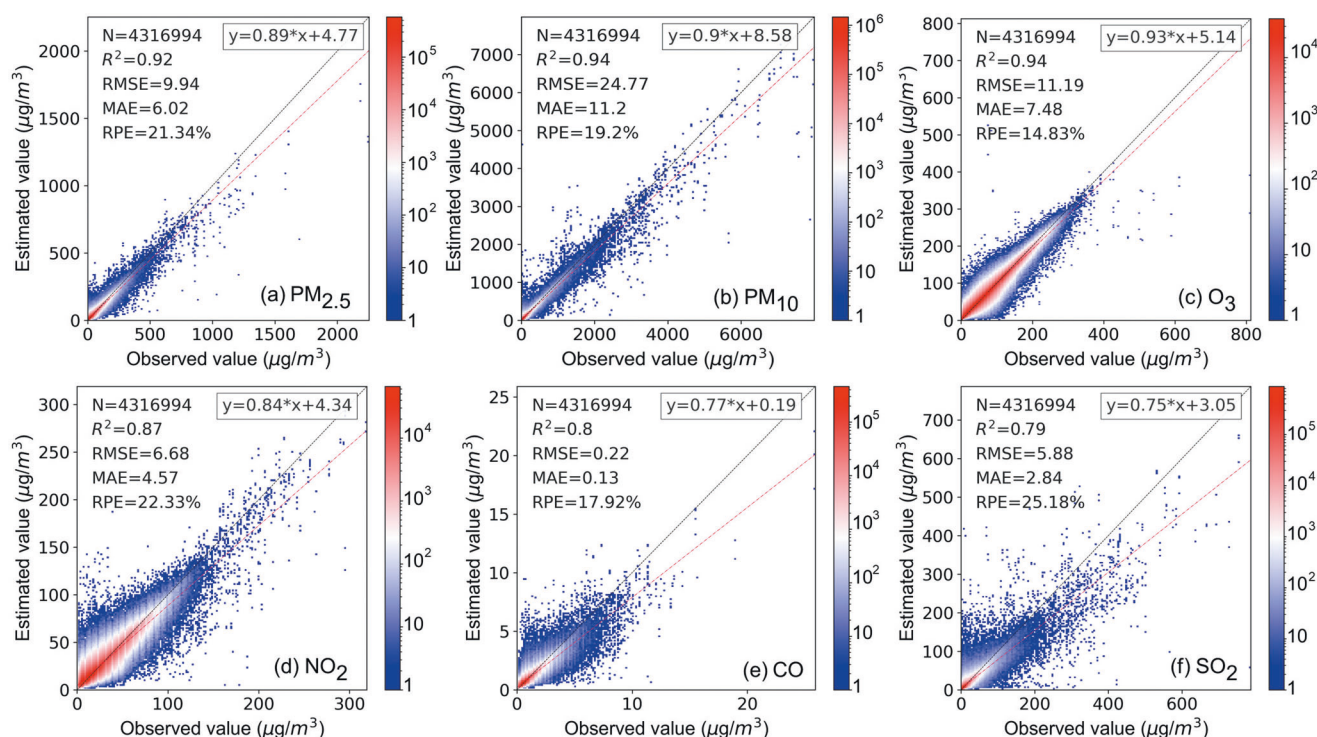
This study aims to achieve a simultaneous estimation of hourly ground-level concentrations of six criteria air pollutants in China using the multi-output random forest model (MORF) and the latest GEMS data. The joint inversion of multiple pollutants proposed here is expected to simplify the process of retrieving surface concentrations of six criteria air pollutants, and largely improve

modeling efficiency compared to traditional models. The proposed method can infer hourly variations of air pollutants with high accuracy and high efficiency, and assist in monitoring the evolution of pollution episodes.

## RESULTS AND DISCUSSION

### Statistics of model performance

Figure 1 displays the results of sample-based CV for hourly retrievals. MORF achieves a general CV  $R^2$  of 0.95 and RPE of 20.13% for six air pollutants, yet the performances vary with pollutant types.  $R^2$  values range from 0.79 to 0.94 and RPE values range from 14.83 to 25.18%. The best performance is yielded for the estimation of  $O_3$  concentrations, with CV  $R^2$ , RMSE, MAE, and RPE of 0.94,  $11.19 \mu\text{g m}^{-3}$ ,  $7.48 \mu\text{g m}^{-3}$ , and 14.83%, respectively. Low bias is also indicated with the slope of the fitting line of 0.93. The model performance for particulates ( $PM_{2.5}$  and  $PM_{10}$ ) also shows high accuracy, which is comparable to that of state-of-the-art single-output models<sup>18,39</sup>. CV  $R^2$  for  $PM_{2.5}$  and  $PM_{10}$  reach 0.92 and 0.94, and the RMSE values are  $9.94 \mu\text{g m}^{-3}$  and  $24.77 \mu\text{g m}^{-3}$ , respectively. The retrieval accuracy for  $NO_2$  and CO are relatively lower, with  $R^2$  of 0.87 and 0.80, and RPE of 22.33 and 17.92%, respectively. MORF model yields the worst performance for  $SO_2$  estimation, with a CV  $R^2$  of 0.79 and RPE of 25.18%. UV-based satellite retrieval of  $SO_2$  has been reported to be subject to large uncertainties due to the presence of  $O_3$  absorption and strong molecular Rayleigh scattering<sup>40</sup>. This might also explain the relatively poorer performance of  $SO_2$  from our approach. The performance of sample-based CV is relatively stable across different hours, months, and stations (Supplementary Note 1). Generally, model performance is relatively better in the warm season for  $O_3$  estimation and in the cold season for other pollutants. The model yields a higher accuracy at noon than in the morning and afternoon. Besides, model performance in regions with limited sites is poorer than that in regions with a large number of ground stations, which are consistent with previous studies<sup>16</sup>.



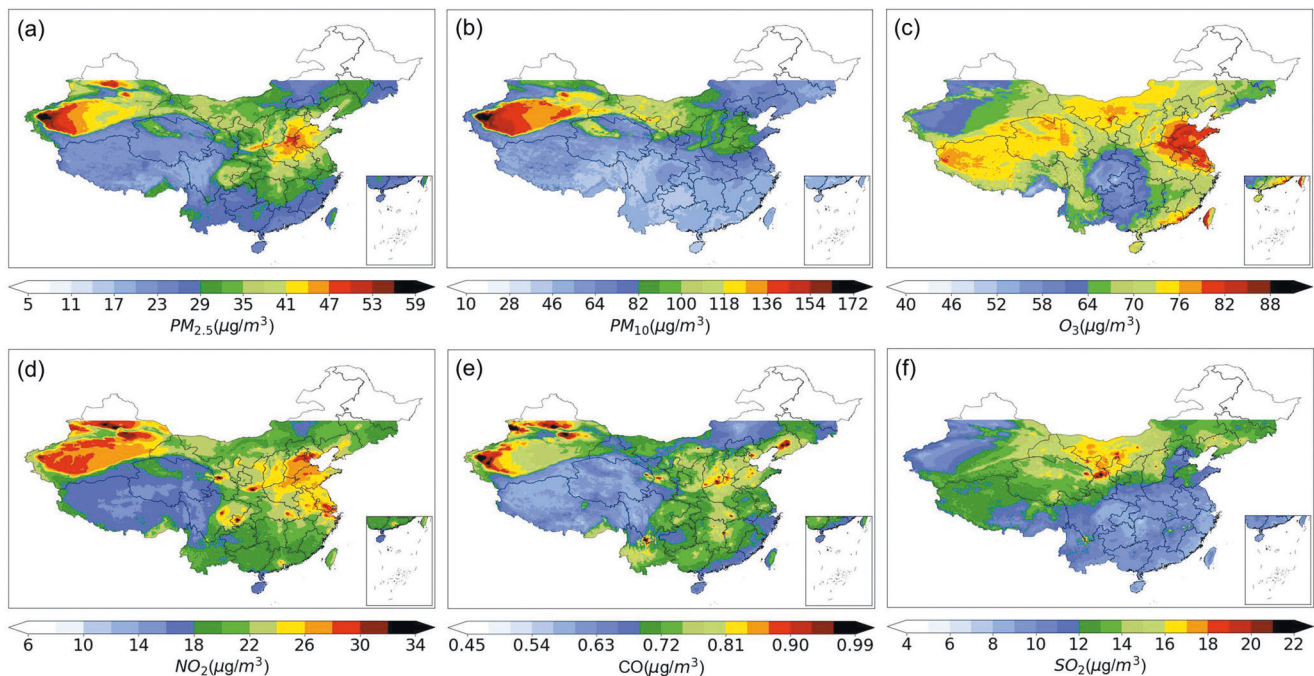
**Fig. 1** Scatter plots of the sample-based tenfold cross-validation results of MORF. **a** For  $PM_{2.5}$ , **b** for  $PM_{10}$ , **c** for  $O_3$ , **d** for  $NO_2$ , **e** for CO, and **f** for  $SO_2$ . The color of the points represents point density, and N means sample number. Dark lines are the 1:1 line and the red dotted lines are the fitted lines.

**Table 1.** Comparisons between the model performance of MORF and SORF.

		Model accuracy						Model efficiency <sup>①</sup>	
		PM <sub>2.5</sub>	PM <sub>10</sub>	O <sub>3</sub>	NO <sub>2</sub>	CO	SO <sub>2</sub>		
MORF	R <sup>2</sup>	0.920	0.938	0.943	0.866	0.802	0.794	Training time (min)	10.46
	RMSE	9.939	24.767	11.194	6.677	0.221	5.884	Prediction time <sup>②</sup> (s)	4.52
	MAE	6.018	11.203	7.481	4.572	0.135	2.844	Model size (GB)	8.62
	RPE(%)	21.337	19.204	14.831	22.329	17.921	25.179		
SORF	R <sup>2</sup>	0.911	0.942	0.943	0.878	0.805	0.793	Training time (min)	48.60
	RMSE	9.906	25.227	11.226	6.615	0.219	5.884	Prediction time (s)	6.84
	MAE	5.970	11.305	7.430	4.484	0.132	2.831	Model size (GB)	31.47
	RPE(%)	21.214	19.312	14.686	22.017	17.611	25.084		

<sup>①</sup> All models were trained on the machine with Intel(R) Core(TM) i7-10700 F CPU @ 2.90 GHz and 16 G RAM.

<sup>②</sup> Prediction time means the time for retrieving one regridded GEMS image.



**Fig. 2** Spatial distributions of annual mean concentrations for six air pollutants. **a** For PM<sub>2.5</sub>, **b** for PM<sub>10</sub>, **c** for O<sub>3</sub>, **d** for NO<sub>2</sub>, **e** for CO, **f** for SO<sub>2</sub>.

Site-based CV results are slightly worse than those of sample-based CV (Supplementary Note 2). R<sup>2</sup> range from 0.56 to 0.91 for different kinds of pollutants. O<sub>3</sub> estimation yields the best accuracy with R<sup>2</sup> of 0.91, RMSE of 13.91  $\mu\text{g m}^{-3}$ , MAE of 9.67  $\mu\text{g m}^{-3}$ , and RPE of 19.20%. For SO<sub>2</sub> and CO, R<sup>2</sup> decreases by ~0.23 and RPE increases by ~8% compared to sample-based CV. Site-based CV R<sup>2</sup> for other pollutants range from 0.75 to 0.84, and RPE from 24.61 to 29.24%.

In addition to CV, we also conducted an independent validation (IV). The results are provided in Supplementary Note 3. The results of IV are similar to that of CV, proving that the proposed model is stable and generalized.

The model performance of MORF was compared with that of SORF in terms of accuracy and efficiency. We trained six separate SORF models, each using one of the six air pollutants as output. The model parameters were the same as the MORF model. The comparison results are listed in Table 1. The retrieval accuracy of MORF and SORF are very close, but MORF outperforms SORF in terms of efficiency. The training of MORF (time for fitting MORF model with all samples) took only 10 min while training six SORF models cost nearly 50 minutes. In addition, MORF took 4.52 s for

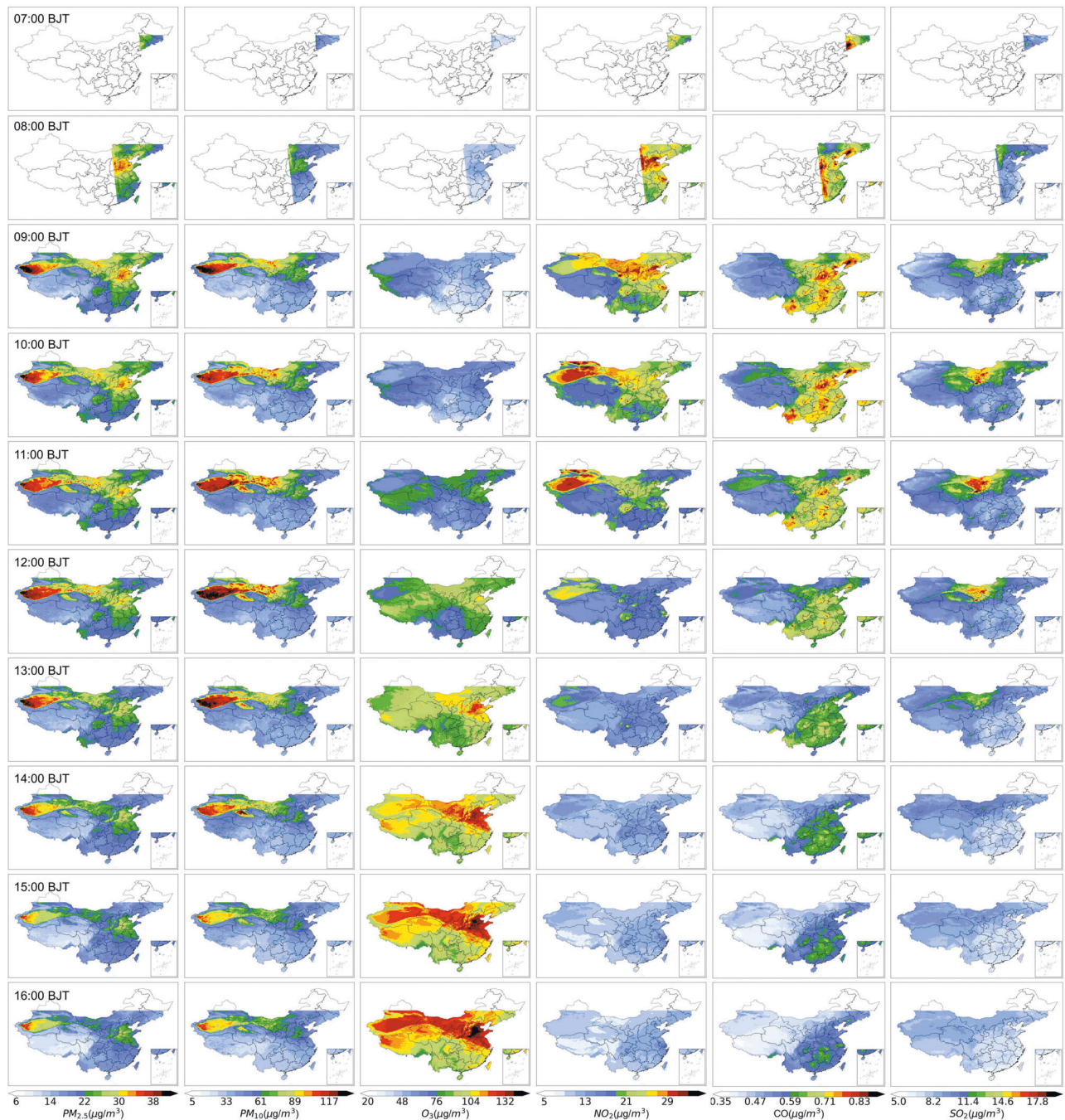
retrieving one resampled GEMS image, but SORF models needed 6.84 s to complete the estimation of six air pollutants. The model size of SORF was also much larger than that of MORF. Considering that building six SORF models also means more efforts on data preparation, data preprocessing, parameter tuning, etc., MORF is much more efficient than SORF.

#### Spatiotemporal variations of six criteria air pollutants

Considering the uneven distribution of GEMS data in different months and hours (Supplementary Note 4), we calculated the monthly mean first and then used the monthly mean values to calculate the annual mean to reduce the bias caused by uneven sample distribution. Besides, we divided data into two parts when analyzing diurnal variation. For the warm season, data for all the hours were considered, while only data from 00:45 UTC to 06:45 UTC were analyzed for cold season.

Spatial distributions of air pollutants in 2021 are displayed in Fig. 2. In terms of spatial variation, PM<sub>2.5</sub> hotspots are located in the junction of Henan, Hebei, and Shandong provinces, and the west of



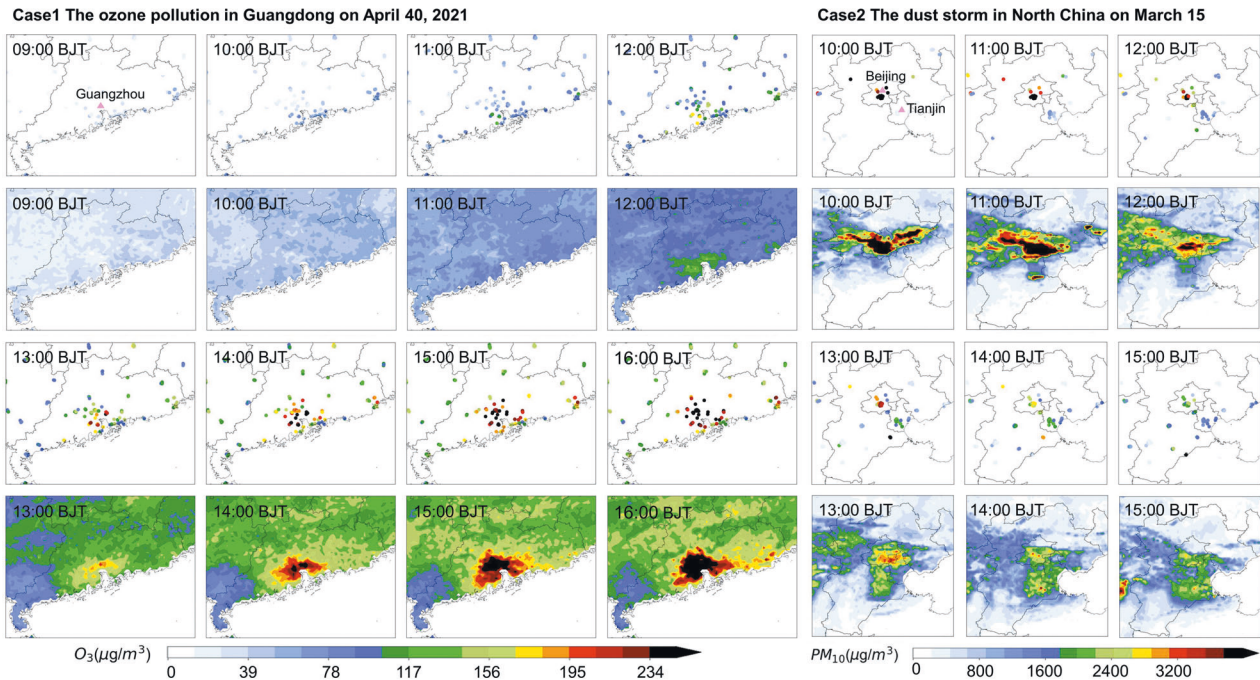


**Fig. 3** Diurnal variations of air pollution in the warm season. Six columns represent diurnal variations of  $PM_{2.5}$ ,  $PM_{10}$ ,  $O_3$ ,  $NO_2$ ,  $CO$ , and  $SO_2$ , respectively. BJT means Beijing Time, and BJT = UTC + 8 h.

Xinjiang (locations of these provinces can be found in Supplementary Note 5). Areas with high  $PM_{10}$  concentration are mainly located in northwestern China, where dust storm happens frequently<sup>41</sup>.  $O_3$  pollution is most serious in Shandong province and surrounding regions and some coastal cities in southern China. The distribution of areas with high  $NO_2$  concentrations is highly consistent with locations of a metropolis, such as the Beijing-Tianjin-Hebei (BTH) region, Yangtze River Delta (YRD), Guangzhou, Wuhan, Chengdu, Chongqing, Lanzhou, and Xian. This is related to its dominant source of transportation<sup>42</sup>. Unlike particulates and  $O_3$  pollution,  $CO$  and  $SO_2$  are more associated with point sources<sup>43</sup>, as indicated by  $CO$  hotspots in Shenyang (Liaoning province), Jincheng (Shanxi), Tangshan (Hebei), Wuhan (Hubei), Lanzhou (Gansu), Xining (Qinghai),

the border of Chuxiong (Yunnan) and Panzhihua (Sichuan), and Xinjiang. The distribution of  $SO_2$  hotspots is similar to that of  $CO$ . Highest  $SO_2$  concentrations are detected in Lanzhou (Gansu), Xining (Qinghai), and some cities in Inner Mongolia, consistent with ground-level observations (Supplementary Note 6). Under national regulations of  $SO_2$  emissions in eastern and southern China,  $SO_2$  concentrations in the YRD, BTH, and PRD have decreased remarkably over recent years. However, in northwestern China,  $SO_2$  concentration keeps growing due to the expansion and relocation of the energy industry<sup>44</sup>. The seasonal variations are consistent with previous studies (details are provided in Supplementary Note 7).

Figure 3 displays the diurnal variations in the warm season.  $PM_{2.5}$  and  $PM_{10}$  concentrations decrease with time in most



**Fig. 4 Variations of pollutant concentrations during two pollution events.** Row 1 and 3 show the measurements from ground stations and row 2 and 4 represent our retrieval results.

regions, which is associated with the development of BLH<sup>45</sup> and the high emissions during morning rush hours. In contrast, in northwestern China, particulate concentrations increase first from 08:00 BJT to 12:00 BJT, and then decrease from 12:00 BJT to 16:00 BJT<sup>46</sup>. Different diurnal variation patterns in northwestern China and other regions can be attributed to the difference in pollution sources. O<sub>3</sub> concentrations increase from 09:00 BJT to 15:00 BJT, due to the enhanced solar radiation and photochemical reaction activity during daytime<sup>22</sup>. Similar to the diurnal pattern of PM<sub>2.5</sub>, NO<sub>2</sub> and CO concentrations decrease from 08:00 BJT to 16:00 BJT gradually under the influence of boundary layer mixing. SO<sub>2</sub> concentrations in northeast China present a decreasing trend during the daytime. However, in Inner Mongolia and northwestern China, SO<sub>2</sub> increases from 8:00 BJT to 11:00 BJT and then decreases.

The diurnal variations in the cold season are basically consistent with those in the warm season, with several small differences (Supplementary Note 8). Particulate concentrations in northwestern China peak in the later noon (14:00 BJ time) rather than at noon (12:00 BJ time) in the cold season. In the warm season, we find the most distinct increase of O<sub>3</sub> concentration happens in the BTH region, which is different from that in the cold season that occurs in southern China. This is related to the different seasonality features of O<sub>3</sub> across China<sup>47</sup>.

### Application in monitoring pollution episodes

We selected two pollution cases to show some examples of how our results can help with monitoring dynamic evolution. As shown in Fig. 4, we use hourly estimations to monitor the dynamic evolution of a serious O<sub>3</sub> pollution event in Guangdong province on April 30, 2021, and a dust storm event in northern China on March 15, 2021. Comparisons with ground-level observations suggest that our retrieved maps accurately capture the changes in O<sub>3</sub> concentrations during this pollution episode. O<sub>3</sub> concentrations increase rapidly from 20 µg m<sup>-3</sup> at 9:00 BJT to >250 µg m<sup>-3</sup> at 16:00 BJT in Guangzhou and surrounding cities. Another small hotspot located in the southeastern corner of Guangdong

province is also detected, where O<sub>3</sub> concentration reaches 200 µg m<sup>-3</sup> at 16:00 BJT. For the dust storm event, ground-level observations indicate an extremely high PM<sub>10</sub> concentration (>3500 µg m<sup>-3</sup>) in Beijing which is also well reflected in the retrieved maps. Besides, both station observations and our retrievals show that PM<sub>10</sub> concentrations in Beijing decrease from >3500 µg m<sup>-3</sup> at 10:00 BJT to ~2500 µg m<sup>-3</sup> at 15:00 BJT. These two cases demonstrate that retrieval results from our proposed algorithm can well capture changes in pollutant concentrations during pollution events.

### DISCUSSIONS

Geostationary satellites offer great potential to monitor air pollution due to their advantage in spatial and temporal coverage. Previously, a number of machine learning models were built to infer ground-level concentrations of air pollutants from satellite images. High estimation accuracy was achieved in these models, yet a joint inversion model that improves modeling efficiency and reduces modeling complexity is still lacking. In this study, we approximated it to a multi-output problem and proposed a unified retrieval model based on MORF that achieved simultaneous estimation of hourly concentrations of six criteria air pollutants in China, benefiting from the world's first geostationary air pollution monitoring spectrometer GEMS. CV results for all samples, different months, hours, and stations demonstrated the accuracy and stability of our MORF model. Comparisons with SORF proved that MORF was much more efficient than a current single-output model. Based on our retrieval results, the spatial, seasonal, and diurnal variations of the six pollutants were analyzed in detail. In general, the maximum values of daytime PM<sub>2.5</sub>, NO<sub>2</sub>, and CO appear in the morning. PM<sub>10</sub> concentrations peak at noon in the warm season and in the afternoon in the cold season. O<sub>3</sub> concentrations increase from morning to afternoon, associated with photochemistry intensity. We also used retrieved maps to monitor the dynamic evolution of pollutants during two pollution events, an O<sub>3</sub> pollution event in Guangdong province and a dust storm event in northern China. Our retrieval results captured the



same variations of pollution as ground stations, but showed better spatial coverage.

Even so, limitations still exist. For instance, the model accuracy of the estimation of SO<sub>2</sub> and CO can be further improved. On the one hand, the absorption features of SO<sub>2</sub> and CO, namely ultraviolet-B and infrared bands, are outside the wavelength range used in this study. Collecting data with a wider range of spectral coverage may help with the improvement of model performance. On the other hand, the information satellite can provide about ground-level air pollution can be limited and difficult to extract, other multi-source data such as emissions and point-of-interest information may also benefit the improvement of estimation accuracy. We also noticed that model performance decreased in regions with limited stations. For example, the estimation accuracy of PM<sub>2.5</sub>, PM<sub>10</sub>, and SO<sub>2</sub> was lower in Tibet. This fact should be considered when the retrieval results are used. In the future, data from more stations in these regions can be used when available to reduce uncertainties<sup>48</sup>. Besides, estimating multiple variables using one model can bring useful extra information, but may also bring mutual interference, especially when uncorrelated tasks are introduced. Therefore, a model that can judge the correlation between multiple tasks may achieve better performance. Some deep-learning-based multi-task models which can evaluate the correlation between different regression tasks and determine the sharing degree according to correlations is worthy of attention. Finally, the physical relationships between ground-level air pollution and satellite radiance data are not fully explained and explored in this study. Interpretable machine learning models can be used in our future work to offer a deeper understanding.

## METHOD

### Data collection

The study area extended from 15°N to 45° N, 73°E to 135°E (Supplementary Material Supplementary Note 5). Ground-level concentrations of the six criteria air pollutants, namely PM<sub>2.5</sub>, PM<sub>10</sub>, SO<sub>2</sub>, NO<sub>2</sub>, CO, and O<sub>3</sub> were obtained from the China National Environmental Monitoring Center website (<http://www.cnemc.cn/en/>). There were more than 1600 stations in 2021. These stations covered all provinces in mainland China and provided pollutant concentrations data with low uncertainty<sup>49</sup>. Hourly data in 2021 were used in this study and negative values were removed as outliers<sup>50</sup>.

Hourly normalized radiance data at six wavelengths (354, 388, 412, 443, 477, and 490 nm), ranging from UV to visible bands, in 2021 were used, which were taken from the GEMS Level 2 (L2) aerosol product<sup>38</sup>. Considering that different air pollutants have different spectral absorption intensities at different wavelengths<sup>51–53</sup>, radiance data at different wavelengths are likely to provide useful information for estimating concentrations of air pollutants. The nominal spatial resolution of the GEMS aerosol product is 3.5 km × 8 km over Seoul, South Korea, and we used hourly data in this study.

The information that satellites images can provide are limited, especially for ground-level trace gases like SO<sub>2</sub> and CO. Therefore, meteorological and spatiotemporal information were also considered in our model. Four meteorological variables, including hourly boundary layer height (BLH), 2 m temperature (T), 2 m dew point temperature (DT), and surface solar radiation downwards (SR), were taken from the ECMWF (European Center for Medium-Range Weather Forecast) Reanalysis v5 (ERA5) dataset<sup>54</sup>. The spatial resolution of BLH was 0.25°×0.25°, while that of the other three variables from ERA5-land dataset<sup>55</sup> was 0.1° × 0.1°.

### Data integration

We resampled all the variables to the defined grids of 0.1° × 0.1° using bilinear interpolation<sup>50,56</sup>, and then ground measurements and raster data were collocated according to time and location (longitude and latitude). Hourly GEMS L2 aerosol products were provided at starting time of observation from 22:45 UTC (Universal Coordinated Time) to 7:45 UTC. Considering that GEMS scanned east-west coverage over ~30 min, air pollution, and meteorological data at the hour closest to the starting time were matched with GEMS data. For example, meteorological and air pollution data at 01:00 UTC were matched with GEMS observations that started at 00:45 UTC.

Previous studies indicated that oversampling technique could improve the quality of training samples and promote the model to better learn the relationship between predictors and target variables<sup>20,57,58</sup>. Random oversampling technique<sup>59</sup> was adopted in this study to facilitate better learning. Details about the oversampling strategy are provided in Supplementary Note 9.

### Model development

Spatiotemporal information, satellite observations, and meteorological variables were used to estimate ground-level concentrations of air pollutants, and the model can be expressed as:

$$\begin{aligned} &(PM_{2.5}, PM_{10}, O_3, NO_2, CO, SO_2) \\ &= f(\text{month}, \text{doy}, \text{hour}, \text{RAA}, R_1, R_2, R_3, R_4, R_5, R_6, \text{BLH}, \text{SR}, T, \text{DT}) \end{aligned} \quad (1)$$

in which, month, day (day of the year), and hour are the temporal information, and RAA stands for relative azimuth angle.  $R_1$ – $R_6$  represent normalized radiance at 354, 388, 412, 443, 477, and 490 nm, while BLH, SR,  $T$ , and DT are the four considered meteorological variables.  $f()$  represents the proposed MORF model.

MORF model was developed from the random forest (RF) model<sup>60</sup>. RF model is a widely used decision-tree-based ensemble-learning model. To overcome overfitting, decision trees in RF were trained using only a random subset of training samples with a random subspace of the input features. Individual trees were then formed using a greedy algorithm that involved, at each split node, the generation of several binary split candidates<sup>61</sup>. We used  $Q_m$  and  $n_m$  to represent the data and number of samples at each tree node  $m$ . For each candidate node split  $\theta = (j, t_m)$  that consisted of a feature  $j$  and a threshold  $t_m$ , data were partitioned into two subsets:  $Q_m^{\text{left}}(\theta)$  with  $n_m^{\text{left}}$  samples and  $Q_m^{\text{right}}(\theta)$  with  $n_m^{\text{right}}$  samples. The quality of a candidate split of node  $m$  was then computed using an impurity function  $H()$ :

$$G(Q_m, \theta) = \frac{n_m^{\text{left}}}{n_m} H(Q_m^{\text{left}}(\theta)) + \frac{n_m^{\text{right}}}{n_m} H(Q_m^{\text{right}}(\theta)) \quad (2)$$

For a single-output regression task (single-output RF, SORF), the impurity function  $H()$  with an L2 error (mean squared error) can be written as:

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2, \text{ where } \bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y \quad (3)$$

For a multi-output regression problem (multi-output RF, MORF), the splitting criteria were modified to compute the average loss across all  $n_t$  outputs<sup>35</sup>. The impurity function was thus changed to:

$$H'(Q_m) = \frac{1}{n_t} \sum_{i=1}^{n_t} H_i(Q_m) \quad (4)$$

where  $n_t$  is the number of outputs (6 in this study), and  $H'(Q_m)$  is the new impurity function.

Three parameters were tuned in our experiments, i.e., the number of trees ( $n_{\text{estimators}}$ ), the minimum number of samples required for internal node split ( $\text{min\_samples}$ ), and the number of features to make the split decision ( $\text{max\_features}$ ). After a parameter sensitivity test (Supplementary Note 10),  $n_{\text{estimators}}$ ,  $\text{min\_samples}$ , and  $\text{max\_features}$  were set as 30, 3, and 3 for a balance of model accuracy and efficiency.

Variable importance in the MORF model was evaluated with permutation importance<sup>19,62</sup>, which was defined to be the decrease in a model score when a single feature value is randomly shuffled. In general, meteorological variables are the most important, followed by radiance data. But for different air pollutants, the variable importance ranking results are different. The detailed results are provided in Supplementary Note 11.

Model performance was evaluated using tenfold cross-validation (CV)<sup>63</sup> and independent validation (IV). Sample-based CV for all samples, different months, hours, and stations were conducted. In addition, a site-based CV was also conducted. For each round of CV, 10% of stations were selected for testing and the rest for training. For IV, we divided the data into two parts. 70% of the data were used for model fitting, CV, and parameter tuning, and then the fitted model was validated on the remaining 30% of the data<sup>64</sup>. Quantitative metrics, including coefficient of determination ( $R^2$ ), root mean squared error (RMSE), mean absolute error (MAE), and relative predictive

error (RPE), were calculated for each air pollutant<sup>56</sup>:

$$R^2 = \left( \frac{\sum_{i=1}^n (\text{obs}_i - \overline{\text{obs}})(\text{est}_i - \overline{\text{est}})}{\sqrt{\sum_{i=1}^n (\text{obs}_i - \overline{\text{obs}})^2 \sum_{i=1}^n (\text{est}_i - \overline{\text{est}})^2}} \right)^2 \quad (\text{unitless}) \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\text{est}_i - \text{obs}_i)^2}{n}} \quad (\text{unit: } \mu\text{g}/\text{m}^3) \quad (6)$$

$$\text{MAE} = \frac{\sum_{i=1}^n |\text{est}_i - \text{obs}_i|}{n} \quad (\text{unit: } \mu\text{g}/\text{m}^3) \quad (7)$$

$$\text{RPE} = \sqrt{\frac{\sum_{i=1}^n (\text{est}_i - \text{obs}_i)^2}{n}} / \overline{\text{obs}} \quad (\text{unit: } \mu\text{g}/\text{m}^3) \quad (8)$$

where  $n$  is the total number of ground sites and  $i$  represents the  $i$ th sites.  $\text{obs}_i$  and  $\text{est}_i$  represent the observed value and the estimated value at the  $i$ th site, respectively.  $\overline{\text{obs}}$  and  $\overline{\text{est}}$  are the mean values for observed and estimated values at all ground sites. A summary of the flowchart of this study is shown in Fig. 5.

## DATA AVAILABILITY

ERA5 reanalysis dataset is freely available from the Copernicus Climate Change Service (C3S) Climate Data Store. ERA5 hourly data on single levels from 1959 to the present and ERA5-Land hourly data from 1950 to the present are used in this study and are accessible at [https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis\\_era5\\_single\\_levels?tab=overview](https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis_era5_single_levels?tab=overview) and <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview>, respectively. The China National Environmental Monitoring Center data is available at <https://quotsoft.net/air/>. GEMS data used in this study were provided by Yonsei University.

## CODE AVAILABILITY

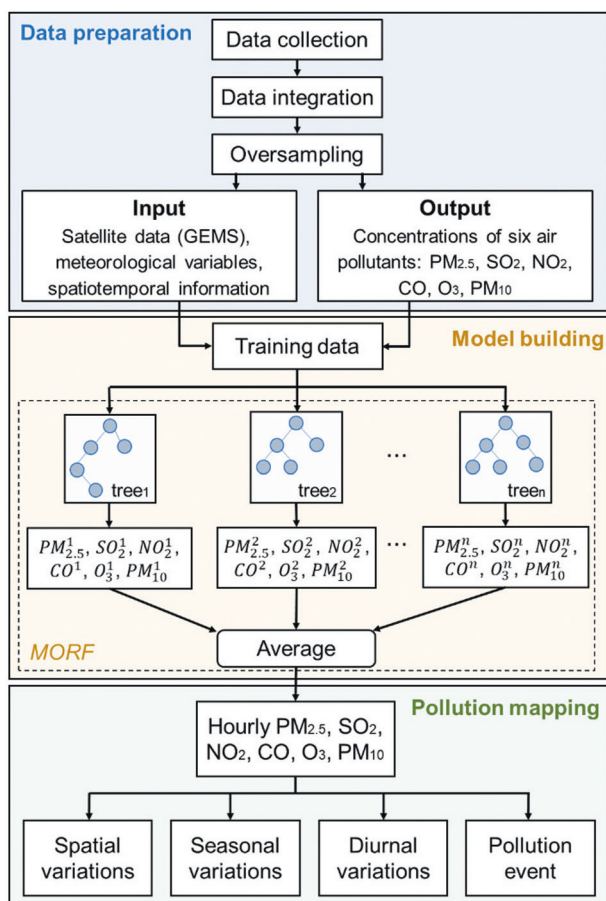
All additional codes needed to perform the analyses are available upon reasonable request from the corresponding author (mmgao2@hkbu.edu.hk).

Received: 23 February 2023; Accepted: 25 June 2023;

Published online: 18 July 2023

## REFERENCES

- Geng, G. et al. Drivers of PM<sub>2.5</sub> air pollution deaths in China 2002–2017. *Nat. Geosci.* **14**, 645–650 (2021).
- Liu, C., Gao, M., Hu, Q., Brasseur, G. P. & Carmichael, G. R. Stereoscopic monitoring: a promising strategy to advance diagnostic and prediction of air pollution. *Bull. Am. Meteorol. Soc.* **102**, E730–E737 (2021).
- Liu, C. et al. Stereoscopic hyperspectral remote sensing of the atmospheric environment: Innovation and prospects. *Earth Sci. Rev.* **226**, 103958 (2022).
- Yang, Q. et al. Mapping PM<sub>2.5</sub> concentration at a sub-km level resolution: a dual-scale retrieval approach. *ISPRS J. Photogramm. Remote Sens.* **165**, 140–151 (2020).
- Kharol, S. K. et al. OMI satellite observations of decadal changes in ground-level sulfur dioxide over North America. *Atmos. Chem. Phys.* **17**, 5921–5929 (2017).
- Cooper, M. J., Martin, R. V., McLinden, C. A. & Brook, J. R. Inferring ground-level nitrogen dioxide concentrations at fine spatial resolution applied to the TROPOMI satellite instrument. *Environ. Res. Lett.* **15**, 104013 (2020).
- Zhang, Y. & Li, Z. Remote sensing of atmospheric fine particulate matter (PM<sub>2.5</sub>) mass concentration near the ground from satellite observation. *Remote Sens. Environ.* **160**, 252–262 (2015).
- Yuan, Q. et al. Deep learning in environmental remote sensing: achievements and challenges. *Remote Sens. Environ.* **241**, 111716 (2020).
- Ma, Z. et al. A review of statistical methods used for developing large-scale and long-term PM<sub>2.5</sub> models from satellite data. *Remote Sens. Environ.* **269**, 112827 (2022).
- Zhang, Y. et al. Satellite remote sensing of atmospheric particulate matter mass concentration: advances, challenges, and perspectives. *Fundam. Res.* **1**, 240–258 (2021).
- Gao, M. et al. Seasonal prediction of Indian wintertime aerosol pollution using the ocean memory effect. *Sci. Adv.* **5**, eaav4157 (2019).



**Fig. 5** The flowchart of this study. Three boxes with different colors represent the three major steps: data preparation, model building, and pollution mapping.

12. Liang, F. et al. Evaluation of a data fusion approach to estimate daily PM<sub>2.5</sub> levels in North China. *Environ. Res.* **158**, 54–60 (2017).
13. He, Q. & Huang, B. Satellite-based high-resolution PM<sub>2.5</sub> estimation over the Beijing-Tianjin-Hebei region of China using an improved geographically and temporally weighted regression model. *Environ. Pollut.* **236**, 1027–1037 (2018).
14. Wang, J. & Christopher, S. A. Intercomparison between satellite-derived aerosol optical thickness and PM<sub>2.5</sub> mass: Implications for air quality studies. *Geophys. Res. Lett.* **30**, 2095 (2003).
15. Liu, Y., Paciorek Christopher, J. & Koutrakis, P. Estimating regional spatial and temporal variability of PM<sub>2.5</sub> concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.* **117**, 886–892 (2009).
16. He, Q. & Huang, B. Satellite-based mapping of daily high-resolution ground PM<sub>2.5</sub> in China via space-time regression modeling. *Remote Sens. Environ.* **206**, 72–83 (2018).
17. Li, T., Shen, H., Yuan, Q., Zhang, X. & Zhang, L. Estimating ground-level PM<sub>2.5</sub> by fusing satellite and station observations: a geo-intelligent deep learning approach. *Geophys. Res. Lett.* **44**, 985–911,993 (2017). 11.
18. Wang, B. et al. Estimate hourly PM<sub>2.5</sub> concentrations from Himawari-8 TOA reflectance directly using geo-intelligent long short-term memory network. *Environ. Pollut.* **271**, 116327 (2021).
19. Yang, N., Shi, H., Tang, H. & Yang, X. Geographical and temporal encoding for improving the estimation of PM<sub>2.5</sub> concentrations in China using end-to-end gradient boosting. *Remote Sens. Environ.* **269**, 112828 (2022).
20. Geng, G. et al. Tracking air pollution in China: near real-time PM<sub>2.5</sub> retrievals from multisource data fusion. *Environ. Sci. Technol.* **55**, 12106–12115 (2021).
21. Wei, J. et al. Full-coverage mapping and spatiotemporal variations of ground-level ozone (O<sub>3</sub>) pollution from 2013 to 2020 across China. *Remote Sens. Environ.* **270**, 112775 (2022).
22. Wang, Y., Yuan, Q., Li, T., Zhu, L. & Zhang, L. Estimating daily full-coverage near surface O<sub>3</sub>, CO, and NO<sub>2</sub> concentrations at a high spatial resolution over China based on SSP-TROPOMI and GEOS-FP. *ISPRS J. Photogramm. Remote Sens.* **175**, 311–325 (2021).
23. Wang, Y., Yuan, Q., Li, T. & Zhu, L. Global spatiotemporal estimation of daily high-resolution surface carbon monoxide concentrations using Deep Forest. *J. Clean. Prod.* **350**, 131500 (2022).
24. Shen, H., Li, T., Yuan, Q. & Zhang, L. Estimating regional ground-level PM<sub>2.5</sub> directly from satellite top-of-atmosphere reflectance using deep belief networks. *J. Geophys. Res. Atmos.* **123**, 13875–13886 (2018).
25. Chen, B. et al. Estimation of atmospheric PM<sub>10</sub> concentration in China using an interpretable deep learning model and top-of-the-atmosphere reflectance data from China's new generation geostationary meteorological satellite, FY-4A. *J. Geophys. Res. Atmos.* **127**, e2021JD036393 (2022).
26. Luo, N. et al. Explainable and spatial dependence deep learning model for satellite-based O<sub>3</sub> monitoring in China. *Atmos. Environ.* **290**, 119370 (2022).
27. Chen, B. et al. Estimation of near-surface ozone concentration and analysis of main weather situation in China based on machine learning model and Himawari-8 TOAR data. *Sci. Total Environ.* **864**, 160928 (2023).
28. Li, M., Yang, Q., Yuan, Q. & Zhu, L. Estimation of high spatial resolution ground-level ozone concentrations based on Landsat 8 TIR bands with deep forest model. *Chemosphere* **301**, 134817 (2022).
29. Yang, Q., Yuan, Q. & Li, T. Ultrahigh-resolution PM<sub>2.5</sub> estimation from top-of-atmosphere reflectance with machine learning: Theories, methods, and applications. *Environ. Pollut.* **306**, 119347 (2022).
30. Gao, M., Ji, D., Liang, F. & Liu, Y. Attribution of aerosol direct radiative forcing in China and India to emitting sectors. *Atmos. Environ.* **190**, 35–42 (2018).
31. Borchani, H., Varando, G., Bielza, C. & Larrañaga, P. A survey on multi-output regression. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **5**, 216–233 (2015).
32. Mandal, D. et al. Crop biophysical parameter retrieval from Sentinel-1 SAR data with a multi-target inversion of Water Cloud Model. *Int. J. Remote Sens.* **41**, 5503–5524 (2020).
33. Tuia, D., Verrelst, J., Alonso, L., Perez-Cruz, F. & Camps-Valls, G. Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geosci. Remote Sens. Lett.* **8**, 804–808 (2011).
34. Bediaga, H. et al. Multi-output chemometrics model for gasoline compounding. *Fuel* **310**, 122274 (2022).
35. Dapogny, A., Bailly, K. & Dubuisson, S. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* 135–140 (2017).
36. Talavera-Llames, R., Pérez-Chacón, R., Troncoso, A. & Martínez-Alvarez, F. MV-kWNN: A novel multivariate and multi-output weighted nearest neighbours algorithm for big data time series forecasting. *Neurocomputing* **353**, 56–73 (2019).
37. Saide, P. E. et al. Assimilation of next generation geostationary aerosol optical depth retrievals to improve air quality simulations. *Geophys. Res. Lett.* **41**, 9188–9196 (2014).
38. Kim, J. et al. New era of air quality monitoring from space: geostationary environment monitoring spectrometer (GEMS). *Bull. Am. Meteorol. Soc.* **101**, E1–E22 (2020).
39. Mao, F. et al. Estimating hourly full-coverage PM<sub>2.5</sub> over China based on TOA reflectance data from the Fengyun-4A satellite. *Environ. Pollut.* **270**, 116119 (2020).
40. Gonzalez Abad, G. et al. Five decades observing Earth's atmospheric trace gases using ultraviolet and visible backscatter solar radiation from space. *J. Quant. Spectrosc. Radiat. Transf.* **238**, 106478 (2019).
41. Li, J. et al. Mixing of Asian mineral dust with anthropogenic pollutants over East Asia: a model case study of a super-duststorm in March 2010. *Atmos. Chem. Phys.* **12**, 7591–7607 (2012).
42. Liu, F. et al. Recent reduction in NO<sub>x</sub> emissions over China: synthesis of satellite observations and emission inventories. *Environ. Res. Lett.* **11**, 114002 (2016).
43. Li, S. & Xie, S. Spatial distribution and source analysis of SO<sub>2</sub> concentration in Urumqi. *Int. J. Hydrog. Energy* **41**, 15899–15908 (2016).
44. Ling, Z. et al. OMI-measured increasing SO<sub>2</sub> emissions due to energy industry expansion and relocation in northwestern China. *Atmos. Chem. Phys.* **17**, 9115–9131 (2017).
45. Gao, M. et al. Reduced light absorption of black carbon (BC) and its influence on BC-boundary-layer interactions during "APEC Blue". *Atmos. Chem. Phys.* **21**, 11405–11421 (2021).
46. Liu, Z. et al. Seasonal and diurnal variation in particulate matter (PM<sub>10</sub> and PM<sub>2.5</sub>) at an urban site of Beijing: analyses from a 9-year study. *Environ. Sci. Pollut. Res. Int.* **22**, 627–642 (2015).
47. Gao, M. et al. Ozone pollution over China and India: seasonality and sources. *Atmos. Chem. Phys.* **20**, 4399–4414 (2020).
48. Zeng, Z. et al. Estimating hourly surface PM<sub>2.5</sub> concentrations across China from high-density meteorological observations by machine learning. *Atmos. Res.* **254**, 105516 (2021).
49. Li, T., Shen, H., Zeng, C., Yuan, Q. & Zhang, L. Point-surface fusion of station measurements and satellite observations for mapping PM<sub>2.5</sub> distribution in China: methods and assessment. *Atmos. Environ.* **152**, 477–489 (2017).
50. Zhou, C. et al. Optimal planning of air quality-monitoring sites for better depiction of PM<sub>2.5</sub> pollution across China. *ACS Environ. Au.* **2**, 314–323 (2022).
51. Krotkov, N. A., Carn, S. A., Krueger, A. J., Bhartia, P. K. & Kai, Y. Band residual difference algorithm for retrieval of SO<sub>2</sub> from the aura ozone monitoring instrument (OMI). *IEEE Trans. Geosci. Remote Sens.* **44**, 1259–1266 (2006).
52. Veefkind, J. P., Haan, J. F. D., Brinksma, E. J., Kroon, M. & Levelt, P. F. Total ozone from the ozone monitoring instrument (OMI) using the DOAS technique. *IEEE Trans. Geosci. Remote Sens.* **44**, 1239–1244 (2006).
53. van Geffen, J. H. G. M. et al. Improved spectral fitting of nitrogen dioxide from OMI in the 405–465 nm window. *Atmos. Meas. Tech.* **8**, 1685–1699 (2015).
54. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
55. Muñoz-Sabater, J. et al. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
56. Yang, Q., Yuan, Q., Li, T. & Yue, L. Mapping PM<sub>2.5</sub> concentration at high resolution using a cascade random forest based downscaling model: evaluation and application. *J. Clean. Prod.* **277**, 123887 (2020).
57. Vu, B. N. et al. Application of geostationary satellite and high-resolution meteorology data in estimating hourly PM<sub>2.5</sub> levels during the Camp Fire episode in California. *Remote Sens. Environ.* **271**, 112890 (2022).
58. Xiao, Q. et al. Separating emission and meteorological contributions to long-term PM<sub>2.5</sub> trends over eastern China during 2000–2018. *Atmos. Chem. Phys.* **21**, 9475–9496 (2021).
59. Xiao, F. Inference-based naïve bayes: turning naïve bayes cost-sensitive. *IEEE Trans. Knowl. Data Eng.* **25**, 2302–2313 (2013).
60. Zeng, Z. et al. Daily global solar radiation in China estimated from high-density meteorological observations: a random forest model framework. *Earth Space Sci.* **7**, e2019EA001058 (2020).
61. Strobl, C., Malley, J. & Tutz, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **14**, 323–348 (2009).
62. Chen, Y.-W., Medya, S. & Chen, Y.-C. Investigating variable importance in ground-level ozone formation with supervised learning. *Atmos. Environ.* **282**, 119148 (2022).
63. Li, T., Shen, H., Zeng, C. & Yuan, Q. A validation approach considering the uneven distribution of ground stations for satellite-based PM<sub>2.5</sub> estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **13**, 1312–1321 (2020).
64. Xiao, Y., Wang, Y., Yuan, Q., He, J. & Zhang, L. Generating a long-term (2003–2020) hourly 0.25° global PM<sub>2.5</sub> dataset via spatiotemporal downscaling of CAMS with deep learning (DeepCAMS). *Sci. Total Environ.* **848**, 157747 (2022).

## ACKNOWLEDGEMENTS

This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (project no. HKBU12202021 and HKBU22201820) and



the National Natural Science Foundation of China (No. 42005084). The authors are grateful to the GEMS science team for providing GEMS aerosol products and to China National Environmental Monitoring Center for providing ground-level air pollution data.

#### AUTHOR CONTRIBUTIONS

M.G. designed the study, and Q.Y. conducted data analysis with help from J.K., Y.C., W.-J.L., D.-W.L., Q.Y., F.W., C.Z., X.Z., X.X., M.G., Y.G., and G.R.C. Q.Y. and M.G. wrote the paper with inputs from all other authors.

#### COMPETING INTERESTS

The authors declare no competing interests.

#### ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41612-023-00407-1>.

**Correspondence** and requests for materials should be addressed to Meng Gao.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023