Ground-level gaseous pollutants across China: daily seamless mapping and long-term spatiotemporal variations

Jing Wei*¹, Zhanqing Li*¹, Jun Wang², Can Li¹, Pawan Gupta³,⁴, Maureen Cribb³

¹. Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
². Department of Chemical and Biochemical Engineering, Iowa Technology Institute, Center for Global and Regional Environmental Research, University of Iowa, USA
³. STI, Universities Space Research Association (USRA), Huntsville, AL, USA
⁴. NASA Marshall Space Flight Center, Huntsville, AL, USA

* Corresponding authors: zli@atmos.umd.edu; weijing_rs@umd.edu
Abstract

Gaseous pollutants at the ground level seriously threaten the urban air quality environment and public health. There are few estimates of gaseous pollutants that are spatially and temporally resolved and continuous over long periods in China. This study takes advantage of big data and artificial intelligence technologies to generate seamless daily maps of three major pollutant gases, i.e., NO$_2$, SO$_2$, and CO, across China from 2013 to 2020 at a uniform spatial resolution of 10 km. Cross-validation illustrated a high data quality on a daily basis for NO$_2$, SO$_2$, and CO, with mean out-of-bag coefficients of determination (root-mean-square errors) of 0.84 (7.99 μg/m$^3$), 0.84 (10.7 μg/m$^3$), and 0.80 (0.29 mg/m$^3$), respectively. They have experienced significant declines and then recoveries during and after the COVID-19 lockdown associated with changes in anthropogenic emissions in eastern China, while surface CO recovered faster than SO$_2$ and NO$_2$. All gaseous pollutants decreased significantly by 0.23, 2.01, and 49 μg/m$^3$ per year ($p < 0.001$) across China during 2013–2020, especially in three urban agglomerations. The declining rates were larger during 2013–2017 but slowed down in recent years. Both the areas and occurrence probabilities of days exceeding air quality standards also gradually shrank and weakened over time, especially for SO$_2$ and CO, which almost disappeared during 2018–2020, suggesting significant improvements in air quality in China. This reconstructed dataset of surface gaseous pollutants, i.e., ChinaHighNO$_2$, ChinaHighSO$_2$, and ChinaHighCO, will benefit future (especially short-term) air pollution and environmental health-related studies.
1. Introduction

Air pollution has been a major environmental concern, affecting human health, weather and climate (Kan et al., 2012; Kinney, 2008; Li et al., 2017b; Anenberg et al., 2022; Sun et al., 2010; Orellano et al., 2020; Murray et al., 2020), and has thus drawn worldwide attention. The sources of air pollution are complex. They include natural sources such as wildfires and anthropogenic emissions, including pollutants discharged from industrial production (e.g., smoke/dust, sulfur oxides, NOx, and VOCs), hazardous substances released from burning coal during heating seasons (e.g., dust, SO2, and CO), and waste gases (e.g., CO, SO2, and NOx) generated by transportation, especially in big cities.

Among various air pollutants, the followings have been most widely recognized, particulate matters (e.g., PM2.5 and PM10), and gaseous pollutants (e.g., O3, NO2, SO2, and CO, among others). Many countries have built ground-based networks to monitor a variety of conventional pollutants in real-time. China has experienced serious ambient air pollution for a long time, prompting the establishment of a large-scale air quality monitoring network (Mee, 2018). Over the years, much effort has been made to model different species of air pollutants. Many studies on particulate matter (e.g., PM1, PM2.5, and PM10) have been carried out with a focus on China (Ma et al., 2022; Fang et al., 2016; Wei et al., 2019b; Chen et al., 2018; Wei et al., 2021a; Wei et al., 2021b; Li et al., 2017a; Zhang et al., 2018). By contrast, ground-level gaseous pollutants have been much less studied.

The global COVID-19 pandemic has motivated many attempts to estimate surface NO2 concentrations (Tian et al., 2020; Who, 2020) from various satellite-retrieved tropospheric NO2 products, e.g., OMI and TROPOMI, by adopting different statistical regression (Chi et al., 2021; Qin et al., 2017; Zhang et al., 2018) and artificial intelligence (Chi et al., 2022; Dou et al., 2021; Chen et al., 2019; Zhan et al., 2018; Wang et al., 2021; Liu, 2021) models. In contrast, studies on surface SO2 and CO with a focus on whole of China are meager, limited by a lack of tropospheric satellite remote-sensing products and weaker signals (Li et al., 2020; Liu et al., 2019; Wang et al., 2021; Han et al., 2022). Such studies still face more challenges, e.g., satellite data gaps and missing values seriously limit their application and the neglect of the spatial and temporal differences in air pollution in the modeling process. In addition, most previous studies mainly focused on studying a single or a few species during relatively short periods of observation.
As such, here, we aim to reconstruct a long-term daily seamless dataset of three ambient gaseous pollutants (i.e., surface NO₂, SO₂, and CO) in China at a uniform 10 km resolution to study air quality. We adopted the spatiotemporal ensemble-learning model to estimate three surface gaseous pollutants from big data. Using this dataset, the long-term spatiotemporal variations of the three gaseous pollutants and the impacts of implementing environmental protection policies and COVID-19 epidemic are investigated.

2. Materials and methods

2.1 Big data

2.1.1 Ground-based measurements

Major input data employed in the study are hourly routine measurements of the ground-level NO₂, SO₂, and CO concentrations of at approximately 2000 reference-grade ground-based monitoring stations across China from 2013 to 2020. Due to a change in the reference state implemented on 31 August 2018 (Mee, 2018), we first converted the concentrations of the three gaseous pollutants to the uniform standard condition (i.e., 273 K and 1013 hPa) for consistency. Daily values for each air pollutant at each station in each year were then averaged from valid hourly measurements that had undergone additional quality-control checks.

2.1.2 Satellite, reanalysis, and model data

Satellite remote sensing data used here include the daily seamless tropospheric NO₂ products (0.25° × 0.25°) generated by first combining OMI/Aura and Global Ozone Monitoring Experiment–2B retrievals (He et al., 2020), and then gap-filling using CAMS tropospheric NO₂ simulations via machine learning (Wei et al., 2022b), and MODIS monthly NDVI (0.05° × 0.05°), LandScan™ annual population (POP, 1 km) (Bright et al., 2000), and the SRTM digital elevation model (DEM, 90 m). ERA5-Land (0.1° × 0.1°) and ERA5 global reanalysis (0.25° × 0.25°) provided hourly meteorological fields (Muñoz-Sabater et al., 2021; Hersbach et al., 2020). The following eight variables form the reanalysis are employed in our study: 2-m temperature (TEM), precipitation (PRE), evaporation (ET), 10-m u- and v-components of wind (WU and WV), boundary-layer height (BLH), relative humidity (RH), and surface pressure (SP). Besides, model-simulated SO₂ and CO
surface mass concentrations were also included from the MERRA-2 and GEOS-FP global reanalysis every 1 and 3 hours at horizontal resolutions of 0.625° × 0.5° and 0.3125° × 0.25°, respectively. CAMS global reanalysis provided three-hour NO\textsubscript{2} simulations modeled on the earth’s surface (every 3 hours, with a horizontal resolution of 0.75° × 0.75°) (Inness et al., 2019). Monthly 0° × 0.1° anthropogenic emissions, i.e., NO\textsubscript{x}, SO\textsubscript{2}, and CO, were collected from CAMS global emission inventories. Here, for these fine-temporal-resolution variables, all hourly-level simulations in a day were first averaged for each grid to calculate daily means. All variables were aggregated or resampled into a 0.1° × 0.1° resolution for consistency.

2.2 Pollutant gas modelling

The current study is an extension of our previous work related to O\textsubscript{3} (Wei et al., 2022a), aimed at extending the long-term and full-coverage mapping of multi-type ground-level gaseous pollutants at a uniform grid resolution of 0.1° × 0.1° across China. Thus, the developed Space-Time Extra-Tree (STET) model was extended to estimate three additional species of surface pollutant gases, i.e., NO\textsubscript{2}, SO\textsubscript{2}, and CO. The uniqueness of this method is that it considers the autocorrelation and differences in air pollution in space and time, improving the model performance on the basis of ensemble learning.

For the surface NO\textsubscript{2} estimation, the STET was applied to the main input variables of NO\textsubscript{2} column amounts, model-simulated surface NO\textsubscript{2} concentrations, and NO\textsubscript{x} emissions; together with the ancillary input variables of the aforementioned meteorological, land, and population (POP) variables, as denoted in Equation 1. Limited by long-term and high-resolution satellite tropospheric SO\textsubscript{2} and CO products, model-simulated surface SO\textsubscript{2} and CO concentrations and emissions were used as main predictors along with the same auxiliary variables as NO\textsubscript{2} to construct the STET model for separately estimating surface SO\textsubscript{2} (Equation 2) and CO (Equation 3):

\[
\text{NO}_{2(ijt)} \sim f_{\text{STET}}(\text{MNO}_{2(ijt)}, \text{SN}_{2(ijt)}, \text{ENOx}_{ijm}, \text{Meteorology}_{ijt}, \text{NDVI}_{ijm}, \text{DEM}_{ijy}, \text{POP}_{ijy}, P_s, P_t) \tag{1}
\]

\[
\text{SO}_{2(ijt)} \sim f_{\text{STET}}(\text{MSO}_{2(ijt)}, \text{ESO}_{2(ijm)}, \text{Meteorology}_{ijt}, \text{NDVI}_{ijm}, \text{DEM}_{ijy}, \text{POP}_{ijy}, P_s, P_t) \tag{2}
\]

\[
\text{CO}_{ijt} \sim f_{\text{STET}}(\text{MCO}_{ijt}, \text{ECO}_{ijm}, \text{Meteorology}_{ijt}, \text{NDVI}_{ijm}, \text{DEM}_{ijy}, \text{POP}_{ijy}, P_s, P_t) \tag{3}
\]
where NO2(𝑖𝑗𝑡), SO2(𝑖𝑗𝑡), and CO(𝑖𝑗𝑡) indicate daily ground-based NO2, SO2, and CO measurements or estimations at one grid (𝑖, 𝑗) on the 𝑡th day of a year; 𝑀𝑁𝑂2(𝑖𝑗𝑡), 𝑀𝑆𝑂2(𝑖𝑗𝑡), 𝑀𝐶𝑂(𝑖𝑗𝑡), and 𝑆𝑁𝑂2(𝑖𝑗𝑡) indicate daily model-simulated surface NO2, SO2, and CO concentrations and the tropospheric NO2 column amount at one grid (𝑖, 𝑗) on the 𝑡th day of a year; 𝐸𝑁𝑂𝑥(𝑖𝑗𝑚), 𝐸𝑆𝑂2(𝑖𝑗𝑚), and 𝐸𝐶𝑂(𝑖𝑗𝑚) indicate monthly anthropogenic NOx, SO2, and CO emissions at one grid (𝑖, 𝑗) on the 𝑚th month of a year; 𝑀𝑒𝑡𝑒𝑜𝑟𝑜𝑙𝑜𝑔𝑦(𝑖𝑗𝑡) represents each meteorological variable at one grid (𝑖, 𝑗) on the 𝑡th day of a year; 𝐷𝐸𝑀(𝑖𝑗) and 𝑃𝑂𝑃(𝑖𝑗) indicate the population at one grid (𝑖, 𝑗) of a year; 𝑃s and 𝑃t indicate the space and time terms, represented by the longitudes and latitudes of spatial points and their distances to the center (𝐷𝑠) and each corner of the study domain, and the day of the year (𝐷OY), respectively.

Here, two widely-used 10-fold out-of-sample and out-of-station cross-validation (CV) methods (Wei et al., 2022a; Rodriguez et al., 2010) were employed to assess the data quality. They were performed by randomly dividing data samples and ground monitoring stations into independent training and testing datasets to evaluate the overall accuracy and prediction reliability, i.e., estimates for the samples and predictions for the stations that are excluded from training, respectively. Wei et al. (2022a) provides details about how these two methods work.

3. Results and discussion

3.1 Model performance

Using the constructed STET models, we generated daily 10-km resolution dataset with complete coverage (spatial coverage = 100%) for three ground-level gaseous pollutants from 2013 to 2020 in China, called ChinaHighNO2, ChinaHighSO2, and ChinaHighCO, respectively. They are all assembled to the ChinaHighAirPollutants (CHAP) dataset. Figure 1 shows the cross-validation results of all daily estimates and predictions for ground-level NO2, SO2, and CO concentrations in China (sample size: N ≈ 3.6 million). Surface NO2 and SO2 concentrations fell between 200 and 500 µg/m3, respectively, and daily estimates were highly correlated to observations, with the same coefficient of determination (R2 = 0.84) and slopes (0.86 and 0.84) close to 1, respectively. Average root-mean-square errors (RMSEs) of surface NO2 and SO2 estimates were 7.99 and 10.07 µg/m3, respectively, and mean absolute errors (MAEs) were 5.34 and 4.68 µg/m3, respectively. Most
observed CO concentrations were less than 10 mg/m$^3$, agreeing well ($R^2 = 0.80$, slope = 0.79) with
daily estimates, and the average RMSE (MAE) is 0.29 (0.16) mg/m$^3$. Compared to estimation
accuracies (Figure 1a-c), prediction accuracies slightly decreased, which is acceptable considering
the weak signals of trace gases. Daily SO$_2$, NO$_2$, and CO predictions (Figure 1d-f) show reasonable
agreements with ground measurements ($R^2 = 0.70$, 0.68, and 0.61, respectively) and their respective
RMSE (MAE) values were 14.28 (8.10) µg/m$^3$, 11.57 (7.06) µg/m$^3$, and 0.42 (0.24) mg/m$^3$, which
basically represent the accuracy for areas without ground monitoring stations.

The performance of our air pollution modeling was also evaluated on an annual basis (Table 2). Our
model works well in estimating and predicting the concentrations of different ground-level pollutant
gases among different years. The model performance has continuously improved over time, as
indicated by increasing correlations and decreasing uncertainties, because of increasing density of
ground stations (especially in the suburban areas of cities) and improved quality control of
measurements, significantly increasing the number (e.g., from 169 thousand in 2013 to more than
522 thousand in 2020) and quality of data samples.

Figure 2 shows the individual-site-scale cross-validated accuracy and uncertainty in estimating
daily pollutant gases in China. Our model has a strong ability to capture daily surface NO$_2$
concentrations at most stations in China, with about 80% (83%) of them having CV-R$^2 > 0.7$
(RMSE < 10 µg/m$^3$) with reference to ground measurements. By contrast, the model did not
perform as well in estimating daily surface SO$_2$ and CO concentrations. Nevertheless, regarding
SO$_2$, 80% of the stations had CV-R$^2$ values > 0.6, RMSE values were < 12 µg/m$^3$. Regarding CO,
83% of the stations had CV-R$^2$ values > 0.6, and RMSE values were generally < 0.4 mg/m$^3$. 

[Please insert Figure 1 here]

[Please insert Table 2 here]

[Please insert Figure 2 here]
Figure 3 shows the monthly and yearly composites of ground-level NO2, SO2, and CO concentrations as a function of ground measurements from all monitoring stations in China for the years 2013 to 2020. On the monthly scale, we collected a total of ~119,000 matched samples of the three gaseous pollutants. Accuracies significantly improved, with increasing R² (decreasing RMSE) values of 0.93 (4.41 µg/m³), 0.97 (4.03 µg/m³), and 0.94 (0.13 mg/m³) for NO2, SO2, and CO, respectively. On the annual scale, more than ~10,000 matched samples were collected, showing better agreement with observations (e.g., R² = 0.94, 0.98, and 0.97) and lower uncertainties (e.g., RMSE = 3.06 µg/m³, 2.46 µg/m³, and 0.07 mg/m³) for NO2, SO2, and CO, respectively. These results illustrate the high quality of our dataset for different gaseous pollutants, illustrating its applicability to the investigations of short-term exposure and long-term variations.

[Please insert Figure 3 here]

3.2 Spatiotemporal variations

3.2.1 Spatial coverage and distribution

Figure 4 shows spatial distributions of the three pollutant gases across China on a typical day (1 January 2018). The spatial patterns of these gaseous pollutants are consistent with those observed on the ground, especially in highly polluted areas, e.g., severe surface NO2 pollution in the North China Plain (NCP) and high surface SO2 emissions in Shanxi Province. The unique advantage of our dataset is that it can provide valuable gaseous pollutant information on a daily basis at locations in China where ground measurements are not available. This addresses the major issues of scanning gaps and the numerous missing values in satellite remote sensing retrievals at cloudy locations, e.g., average daily spatial coverage of the OMI tropospheric NO2 product is only 42%. Our dataset provides spatial-complete coverage, significantly increasing the daily data utilization by 58%.

[Please insert Figure 4 here]

Figure 5 shows seasonal maps for each gas pollutant during the period 2013–2020 across China. Pollutant gases vary significantly in space and time across China, where high surface NO2 levels are...
mainly distributed in typical urban agglomerations, e.g., the Beijing-Tianjin-Hebei (BTH) region, the Yangtze River and Pearl River Deltas (YRD and PRD), and scattered large cities with intensive human activities and highly developed transportation systems (e.g., Urumqi, Chengdu, Xi'an, and Wuhan, among others). High surface SO$_2$ concentrations are mainly observed in northern China (e.g., Shanxi, Hebei, and Shandong Provinces), associated with combustion emissions from anthropogenic sources, and the Yunnan Guizhou Plateau in southwest China, likely associated with emissions from volcanic eruptions. By contrast, except in some areas in central China (e.g., Shanxi and Hebei), surface CO concentrations are overall low.

Over time, significant differences in spatial patterns were seen. Surface NO$_2$, SO$_2$, and CO in summer (average = 15.9 ± 4.7 µg/m$^3$, 22.9 ± 13.4 µg/m$^3$, and 1.1 ± 0.3 mg/m$^3$, respectively) were the lowest, thanks to favorable meteorological conditions. Pollution levels were highest in winter, with average values increasing by ~1.5–1.9 times those in summer. This difference was much larger in central and eastern China, e.g., 2.3–3.4 times higher in the BTH, due to large amounts of direct NO$_x$, SO$_2$, and CO emissions from burning coal for heating in winter in northern China. Spring and autumn show relatively similar spatial patterns among three gaseous pollutants.

[Please insert Figure 5 here]

### 3.2.2 Short-term epidemic effects on air quality

The unique advantage of seamless day-to-day gaseous pollutant maps allows us to investigate the COVID-19 effects (Who, 2020) in China. Here, we compared the relative differences of each air pollutant from February to April between 2020 and 2019 in eastern China (Figure 6). In February, surface NO$_2$ sharply reduced by more than 30% in eastern China, especially in key urban agglomerations and megacities (relative change > 50%). A significant decrease in surface SO$_2$ (> 40%) was observed in northern areas where heavy industry is the mainstay in China, while there was little change in southern China. Surface CO also showed drastic decreases, especially in southeast China, but the amplitude was smaller than those of the other two gaseous pollutants. These were attributed to extensive plant closures and traffic controls due to the lockdown, which started at the end of January 2020, significantly reducing anthropogenic NO$_x$, SO$_2$, and CO
emissions (Ding et al., 2020; Zheng et al., 2021).

In March, surface NO$_2$ concentrations were still largely lower than typical levels in most areas, especially in areas hit hardest by the epidemic, i.e., Hubei province and its surrounding areas (relative change > 30%). By comparison, the decline in surface SO$_2$ largely slowed by more than two times in the NCP and Shanxi Province. However, surface CO returned to normal levels the fastest. In April, surface NO$_2$ and SO$_2$ were comparable to historical concentrations (within ±10%) or alternately changed across eastern China due to rebounding anthropogenic emissions (Ding et al., 2020), indicating that they were almost recovered. In addition, the three air pollutants fell within ±10% around Wuhan, Hubei Province, indicating that people there had returned to normal life.

3.2.3 Long-term trends and policy implications

To better investigate the spatiotemporal variations of surface air pollution, we calculated linear trends and significance levels using monthly anomalies by removing seasonal cycles (Wei et al., 2019a). Given that monitoring stations were sparse and unevenly distributed in western China, especially in earlier years, we will focus on eastern part of the country for the trend analysis. Figure 7 shows annual mean maps of the three gaseous pollutants for each year from 2013 to 2020 in eastern China. They all have changed greatly in the past eight years across China, peaking around 2017 and have declined to their lowest levels in 2020, at annual mean surface decreasing rates of -12%, -55%, and -31% across China for NO$_2$, SO$_2$, and CO concentrations, respectively.

Most of eastern China showed more significant decreasing trends, especially in three urban agglomerations (trend = -0.51~1.21 µg/m$^3$/yr, $p < 0.001$), as well as in other large cities (e.g., Wuhan and Chengdu) (Figure 8, and Table 3). The largest downward trends mainly occurred in northern and central China, especially in the BTH (trend = -6.01 and 109 µg/m$^3$/yr, $p < 0.001$, respectively). This is mainly due to the change in fuel for heating from coal to gas widespread
across China in winter (Wang et al., 2020), greatly reducing the emissions of precursor gases
(Koukouli et al., 2018). Increasing trends were, however, also found in Ningxia and Shanxi in
central China.

We divided the study into four periods to investigate the impact of implemented major emission
control measures taken in China. During the Clear Air Action Plan (CAAP, 2013–2017), surface
SO\(_2\) and CO concentrations significantly decreased in most parts of eastern China, while surface
NO\(_2\) reductions are limited to some places. The rates of decreases for all three pollutants accelerated
in recent years, especially for NO\(_2\) in southeast China (trend > 2 µg/m\(^3\)/yr, \(p < 0.05\)). These are
thanks to the dramatic reductions in all main pollutant emissions (e.g., PM, SO\(_2\), and NO\(_x\)) in key
regions (especially urban) through the upgrading of key industries, industrial structure adjustment,
and coal-fired boiler remediation.

During the 13\(^{\text{rd}}\) Five-Year-Plan (FYP, 2016–2020), surface NO\(_2\) decreased at a rate of -0.46 µg/m\(^3\)
per year (\(p < 0.001\)) across China, with larger decreases in the BTH region and Hunan Province
(trend > 2 µg/m\(^3\)/yr, \(p < 0.05\)). More striking decreasing trends were found in southeast China.
Surface SO\(_2\) also significant decreased but slowed down in eastern China. However, a greater
downward trend was observed in Shanxi Province, mainly due to the reduction in coal consumption.
Surface CO also continuous decreased, more rapidly in central China but less rapidly elsewhere.
The continuous decline in all pollutants is due to the binding reduction by 10–15% of total
emissions of major pollutants like COD, SO\(_2\), and NO\(_x\).

During the Blue Sky Defense War (2018–2020), the majority of pollutants have dropped
considerably, of which surface SO\(_2\) changed the most in China (average = -27%), especially in
central and eastern China (relative change > 50%). In addition, SO\(_2\) decreased by 23–44% in three
key regions. Followed by surface CO with the national and regional mean reductions of 17% and
10–24%, respectively. By contrast, surface NO\(_2\) had the least reduction at an average decreasing
rate of 11% across China and 9–16% for the three key regions. The improvement in air quality
benefited from continuous reductions (by 2–3%) in total air pollutant emissions, coordinated

[Please insert Figure 8 and Table 3 here]
With the daily seamless datasets, we can investigate the number of days exceeding their respective air quality standard (Level 2 limitation) in a given year to evaluate the distribution and variations of short-term pollution exposure (Figure 9). The areal extent of regions exposed to unacceptably high pollutant-gas levels (i.e., daily NO$_2$ > 80 μg/m$^3$, SO$_2$ > 150 μg/m$^3$, and CO > 4 mg/m$^3$) were usually small. NO$_2$ pollution was mainly found in the NCP and a handful of big cities (e.g., Xi’an, Wuhan, Guangzhou, and Shanghai), changing little over time. Surface SO$_2$ pollution was mainly observed in central China (e.g., Hebei, Shandong, and Shanxi), with the areal extent of polluted regions gradually decreasing over time until almost disappearing by 2020. The same was seen for surface CO pollution, being worst in the BTH region and its surrounding areas before 2018, then almost disappearing by 2020. Surface NO$_2$ pollution was mainly observed in developed urban areas (e.g., Beijing, Tianjin, Shijiazhuang, and Wuhan), with 15% of the days exceeding the acceptable standard in the early part of the study period, then decreasing to below 5% afterward.

Regionally, significant differences in the percentage of days with pollution levels exceeding national standards are seen (Figure 10). For example, the BTH region was the only region experiencing a high NO$_2$ exposure risk, which gradually lessened from 2013 to 2018, when the exposure risk reached zero. For surface SO$_2$, no high exposure days (daily mean > 150 μg/m$^3$) were observed regionally, but a large number of days exceeding the Level 1 limitation (i.e., daily mean > 50 μg/m$^3$) were found in the BTH region. However, the number of days gradually decreased until reaching zero after 2017. For surface CO, less than 1% of highly polluted days are found in the BTH region in some individual years. These results suggest that with the unrelenting efforts of air pollution prevention and control, the number of days with high pollution has been significantly reduced across China, indicating significant improvements in air quality.
3.3 Model comparison

Long-term datasets (at least 2 years) of different gaseous pollutants generated with different developed models in previous studies focusing on the whole of China are compared here. Only those studies applying the same validation approach were selected (Table 4). Most generated surface NO2 datasets had low spatial resolutions (~0.125°–0.25°) with numerous missing values limited by OMI satellite observations (Zhan et al., 2018; Dou et al., 2021; Chen et al., 2019; Xu et al., 2019; Chi et al., 2021). Some studies improved the spatial resolutions by introducing NO2 data from the recently launched Sentinel-5 TROPOMI satellite but can only provide dataset after October 2018 (Chi et al., 2022; Liu, 2021; Wang et al., 2021; Wei et al., 2022b). Surface SO2 estimated from a SO2 emission inventory and surface CO from the MOPITT and TROPOMI retrievals (Li et al., 2020; Liu et al., 2019; Wang et al., 2021) have a much lower data quality (Li et al., 2020; Li et al., 2017b; Wang et al., 2021). Overall, our gaseous pollutant datasets are superior to those from the studies listed in Table 4 in terms of either overall accuracy, or spatial coverage, or length of data records.

[Please insert Table 4 here]

3.4 Successful applications

Our surface gaseous pollutant datasets have been freely available to the public online since March 2021 and have now been successfully employed for various application studies in environment and health. Strong associations and negative effects between ambient gaseous pollution (e.g., NO2, SO2, and CO) and a variety of diseases has been demonstrated for people of all ages through multi-regional and national cohort studies in China. These diseases include general mortality (Zhang et al., 2022), cause-specific cardiovascular disease (Xu et al., 2022a), ischemic and hemorrhagic stroke (Xu et al., 2022b; Wu et al., 2022b; Cai et al., 2022; He et al., 2022), dementia mortality (Liu et al., 2022), blood pressure (Song et al., 2022; Wu et al., 2022a), renal function (Li et al., 2022a), neurodevelopmental delay (Su et al., 2022), serum liver enzymes (Li et al., 2022b), overweight and obesity (Chen et al., 2022b), insomnia (Xu et al., 2021), subjective sleep quality (Wang et al., 2022), and visual impairment (Chen et al., 2022a). These studies attest well to the value of the
CHAP dataset with some unique merits for the sake of public health, among others, now and in the future.

4. Summary and conclusions

Exposure to air pollution is detrimental to human health, which has been a major public concern in heavy polluted regions like China where ground-based observation of pollutants is relatively short and sparse than major developed countries. Moreover, the pollutants to travel long distances to affect large down stream regions. To remedy the limitations of ground-based air pollution observation, this study applies the machine learning model of Space-Time Extra-Tree to estimate ambient gaseous pollutants across China, with extensive input variables measured on the monitors, satellites, and models. The estimated quantities are daily 10 km resolution (about 0.1 degrees) seamless (spatial coverage = 100%) dataset for ground-level NO$_2$, SO$_2$, and CO concentrations in China since 2013. This dataset was cross-evaluated in terms of overall accuracy and predictive ability at different spatiotemporal levels. Daily estimates (predictions) of surface NO$_2$, SO$_2$, and CO from 2013 to 2020 across China are highly consistent with observations with average sample-based (station-based) CV-R$^2$ of 0.84 (0.68), 0.84 (0.7), and 0.8 (0.61), and RMSEs of 7.99 (11.57) μg/m$^3$, 10.7 (14.28) μg/m$^3$, and 0.29 (0.42) mg/m$^3$/y.

Pollutant gas concentrations varied significantly in the region, where high levels were mainly found in the Northern China, especially in winter. All gaseous pollutants sharply declined in eastern China during the COVID-19 outbreak, then gradually returned to historical levels. The recovery speed of surface CO was faster than for NO$_2$ and SO$_2$. Over time and at the national scale, they have significantly ($p < 0.001$) decreased by 0.23, 2.01, and 43 μg/m$^3$ per year during 2013–2020. Larger reductions were found at the regional scale, especially the BTH. Improvements in air quality were larger in the last decade or so but have slowed down in recent years. In particular, the areal extents of regions experiencing air pollution and the probability of air pollution occurring have also gradually decreased over time, especially during the period 2018–2020. This may be related with the implementation of a series of tough environmental protection policies, which greatly reduced anthropogenic emissions and significantly improved air quality. This high-quality daily seamless dataset will benefit future environmental and health-related studies focused on China, especially...
studies investigating short-term air pollution exposure.

Data availability
CNEMC gaseous pollutants measurements are available at http://www.cnemc.cn; Reconstructed OMI/Aura tropospheric NO₂ product is available at https://doi.org/10.6084/m9.figshare.13126847; MODIS series products and MERRA2 reanalysis are available at https://search.earthdata.nasa.gov/; SRTM DEM is available at https://www2.jpl.nasa.gov/srtm/; LandScan™ population is available at https://landscan.ornl.gov/; ERA5 reanalysis is available at https://cds.climate.copernicus.eu/; GEOS CF data is available at https://portal.nccs.nasa.gov/datashare/gmao/; CAMS reanalysis and emission inventory are available at https://ads.atmosphere.copernicus.eu/

CHAP dataset availability
The ChinaHighAirPollutants (CHAP) dataset is open access and freely available at https://weijing-
rs.github.io/product.html. The ChinaHighNO₂ dataset is available at https://doi.org/10.5281/zenodo.4641542, the ChinaHighSO₂ dataset is available at https://doi.org/10.5281/zenodo.4641538, and the ChinaHighCO dataset is available at https://doi.org/10.5281/zenodo.4641530.

Acknowledgments
JW, and ZL, and JW were supported by the NASA grants (80NSSC21K1980 and 80NSSC19K0950). PG was supported by NASA’s Research Opportunities in Space and Earth Science (ROSES-2020), Program Element A.38: Health and Air Quality Applied Sciences Team.
References


Song, J., Du, P., Yi, W., Wei, J., Fang, J., Pan, R., Zhao, F., Zhang, Y., Xu, Z., Sun, Q., Liu, Y.,


Figures 627

Figure 1. Density plots of daily (a-c) estimates and (d-f) predictions of ground-level NO\(_2\) (µg/m\(^3\)), SO\(_2\) (µg/m\(^3\)), and CO (mg/m\(^3\)) concentrations as a function of ground measurements in China from 2013 to 2020 using the out-of-sample (top panels) and out-of-station (bottom panels) cross-validation methods.
Figure 2. Validation of daily ground-level three NO\textsubscript{2} (µg/m\textsuperscript{3}), SO\textsubscript{2} (µg/m\textsuperscript{3}), and CO (mg/m\textsuperscript{3}) estimates at each individual monitoring station in China from 2013 to 2020: (a-c) accuracy (i.e., $CV - R^2$) and (d-f) uncertainty (i.e., $RMSE$).
Figure 3. Validation results of (a-c) monthly and (d-f) yearly composites of ground-level NO$_2$ ($\mu$g/m$^3$), SO$_2$ ($\mu$g/m$^3$), and CO (mg/m$^3$) against ground measurements from all monitoring stations in China for the years 2013 to 2020. Black lines are best-fit lines from linear regression, and dashed lines are 1:1 lines.
Figure 4. Comparisons between (a-c) big-data-derived (horizontal resolution = 10 km) seamless 
ground-level NO\textsubscript{2} (µg/m\textsuperscript{3}), SO\textsubscript{2} (µg/m\textsuperscript{3}), and CO (mg/m\textsuperscript{3}) concentrations and (d-f) corresponding 
ground measurements on 1 January 2018 in China.
Figure 5. Seasonal mean maps (horizontal resolution = 10 km) of ground-level NO$_2$ (µg/m$^3$), SO$_2$ (µg/m$^3$), and CO (mg/m$^3$) concentrations averaged for 2013–2020 in China.
Figure 6. Relative changes (%) in mean ground-level NO\textsubscript{2}, SO\textsubscript{2}, and CO concentrations (µg/m\textsuperscript{3}) in February, March, and April between 2019 and 2020 during the COVID-19 epidemic across the East China. The area outlined in magenta and the star indicate Hubei Province and Wuhan City, respectively.
Figure 7. Spatial distributions of annual mean (horizontal resolution = 10 km) of ground-level NO$_2$ (µg/m$^3$), SO$_2$ (µg/m$^3$), and CO (mg/m$^3$) concentrations for each year from 2013 to 2020 in China.
Figure 8. Temporal trends (µg/m³/yr) of ground-level NO₂, SO₂, and CO concentrations in eastern China during the whole period (2013–2020), Clean Air Action Plan (2013–2017), 13th Five-Year Plan (2016–2020), where only regions with trends that are significant at the 95% (p < 0.05) confidence level are shown, and relative changes (%) during the Blue Sky Defense War (2018–2020).
Figure 9. Spatial distributions of the percentage of polluted days exceeding air quality standards for ground-level NO\textsubscript{2} (daily mean > 80 $\mu$g/m$^3$), SO\textsubscript{2} (daily mean > 150 $\mu$g/m$^3$), and CO (daily mean > 4 mg/m$^3$) for each year from 2013 to 2020 in eastern China.
Figure 10. Percentage of days exceeding the air quality standards for ground-level (a) NO$_2$ (daily mean $> 80 \mu g/m^3$), (b) SO$_2$ (daily mean $> 150 \mu g/m^3$), (c) SO$_2$ (daily mean $> 50 \mu g/m^3$), and (d) CO (daily mean $> 4 mg/m^3$) for each year from 2013 to 2020 in three typical urban agglomerations in China.
## Table 1. Summary of big data used in this study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Scientific Dataset</th>
<th>Abbreviation</th>
<th>Spatial Resolution</th>
<th>Temporal Resolution</th>
<th>Time Period</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurements</td>
<td>NO₂, SO₂, CO</td>
<td>-</td>
<td>In-situ</td>
<td>Hourly</td>
<td>2013–2020</td>
<td>MEE</td>
</tr>
<tr>
<td>Satellite remote sensing products</td>
<td>Tropospheric NO₂ column</td>
<td>NO₂</td>
<td>0.25° × 0.25°</td>
<td>Daily</td>
<td>2013–2020</td>
<td>(He et al., 2020)</td>
</tr>
<tr>
<td></td>
<td>Normalized difference vegetation index</td>
<td>NDVI</td>
<td>0.05° × 0.05°</td>
<td>Monthly</td>
<td>2013–2020</td>
<td>MOD13C2</td>
</tr>
<tr>
<td></td>
<td>Surface elevation</td>
<td>DEM</td>
<td>90 m</td>
<td>-</td>
<td>-</td>
<td>SRTM</td>
</tr>
<tr>
<td>Population distribution</td>
<td>POP</td>
<td>1 km</td>
<td>Annual</td>
<td>2013–2020</td>
<td>LandScanTM</td>
<td></td>
</tr>
<tr>
<td>Model simulation</td>
<td>2-m air temperature</td>
<td>TEM</td>
<td>0.1° × 0.1°</td>
<td>Hourly</td>
<td>2013–2020</td>
<td>ERA5 reanalysis</td>
</tr>
<tr>
<td></td>
<td>Precipitation</td>
<td>PRE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evaporation</td>
<td>ET</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Surface pressure</td>
<td>SP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10-m u-component of wind</td>
<td>WU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10-m v-component of wind</td>
<td>WV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boundary-layer height</td>
<td>BLH</td>
<td>0.25° × 0.25°</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative humidity</td>
<td>RH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SO₂ surface mass concentration</td>
<td>SO₂</td>
<td>0.3125° × 0.25°</td>
<td>3-hour</td>
<td>2015–2020</td>
<td>GEOS-FP reanalysis</td>
</tr>
<tr>
<td></td>
<td>CO surface concentration</td>
<td>CO</td>
<td>0.625° × 0.5°</td>
<td>Hourly</td>
<td>2013–2020</td>
<td>MERRA2 reanalysis</td>
</tr>
<tr>
<td></td>
<td>NO₂ surface concentration</td>
<td>NO₂</td>
<td>0.75° × 0.75°</td>
<td>3-hour</td>
<td>2013–2020</td>
<td>CAMS reanalysis</td>
</tr>
<tr>
<td></td>
<td>Carbon monoxide</td>
<td>CO</td>
<td>0.1° × 0.1°</td>
<td>Monthly</td>
<td>2013–2020</td>
<td>CAMS emission</td>
</tr>
<tr>
<td></td>
<td>Nitrogen oxides</td>
<td>NOₓ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sulphur dioxide</td>
<td>SO₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Statistics of the overall accuracies and predictive abilities of ambient gaseous pollutants for each year in China from 2013 to 2020.

<table>
<thead>
<tr>
<th>Year</th>
<th>Sample size (N (10^3))</th>
<th>Overall accuracy</th>
<th>Predictive ability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NO₂</td>
<td>SO₂</td>
<td>CO</td>
</tr>
<tr>
<td></td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
</tr>
<tr>
<td>2013</td>
<td>169</td>
<td>0.77</td>
<td>12.48</td>
</tr>
<tr>
<td>2014</td>
<td>324</td>
<td>0.76</td>
<td>10.97</td>
</tr>
<tr>
<td>2015</td>
<td>518</td>
<td>0.79</td>
<td>9.34</td>
</tr>
<tr>
<td>2016</td>
<td>516</td>
<td>0.82</td>
<td>8.59</td>
</tr>
<tr>
<td>2017</td>
<td>527</td>
<td>0.86</td>
<td>7.57</td>
</tr>
<tr>
<td>2018</td>
<td>513</td>
<td>0.87</td>
<td>6.92</td>
</tr>
<tr>
<td>2019</td>
<td>515</td>
<td>0.87</td>
<td>6.78</td>
</tr>
<tr>
<td>2020</td>
<td>522</td>
<td>0.89</td>
<td>5.78</td>
</tr>
</tbody>
</table>
Table 3. Statistics of temporal trends (μg/m³/yr) and relative changes (%) of ground-level NO₂, SO₂, and CO concentrations during the whole period (TAII, 2013–2020), the Clear Air Action Plan (TCAAP, 2013–2017), the 13th Five-Year-Plan (TFYP, 2016–2020), and the Blue Sky Defense War (ΔBSDW, 2018–2020) in China and three typical regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>NO₂</th>
<th>SO₂</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TAI</td>
<td>TCAAP</td>
<td>TFYP</td>
</tr>
<tr>
<td>China</td>
<td>-0.23***</td>
<td>-0.06</td>
<td>-0.46***</td>
</tr>
<tr>
<td>BTH</td>
<td>-1.21***</td>
<td>-1.04***</td>
<td>-1.43***</td>
</tr>
<tr>
<td>YRD</td>
<td>-0.58***</td>
<td>-0.88***</td>
<td>-0.33</td>
</tr>
<tr>
<td>PRD</td>
<td>-0.51***</td>
<td>-0.91**</td>
<td>-0.21</td>
</tr>
</tbody>
</table>

Note: * p < 0.05, ** p < 0.01, and *** p < 0.001.
Table 4. Comparison with long-term datasets of different gaseous pollutants focusing on the whole of China generated in previous related studies.

<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>Missing values</th>
<th>Spatial resolution</th>
<th>Main input</th>
<th>Validation period</th>
<th>CV-R²</th>
<th>RMSE</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO₂</td>
<td>RF-STK</td>
<td>Yes</td>
<td>0.25°</td>
<td>OMI</td>
<td>2013–2016</td>
<td>0.62</td>
<td>13.3</td>
<td>(Zhan et al., 2018)</td>
</tr>
<tr>
<td></td>
<td>RF-K</td>
<td>Yes</td>
<td>0.25°</td>
<td>OMI</td>
<td>2013–2018</td>
<td>0.64</td>
<td>11.4</td>
<td>(Dou et al., 2021)</td>
</tr>
<tr>
<td></td>
<td>KCS</td>
<td>Yes</td>
<td>0.125°</td>
<td>OMI</td>
<td>2014–2016</td>
<td>0.72</td>
<td>7.9</td>
<td>(Chen et al., 2019)</td>
</tr>
<tr>
<td>SO₂</td>
<td>LUR</td>
<td>Yes</td>
<td>0.125°</td>
<td>OMI</td>
<td>2014–2015</td>
<td>0.78</td>
<td>-</td>
<td>(Xu et al., 2019)</td>
</tr>
<tr>
<td></td>
<td>LME</td>
<td>Yes</td>
<td>0.1°</td>
<td>OMI</td>
<td>2014–2020</td>
<td>0.65</td>
<td>7.9</td>
<td>(Chi et al., 2021)</td>
</tr>
<tr>
<td>CO</td>
<td>XGBoost</td>
<td>Yes</td>
<td>0.125°</td>
<td>TROPOMI</td>
<td>2018–2020</td>
<td>0.67</td>
<td>6.4</td>
<td>(Chi et al., 2022)</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>Yes</td>
<td>0.05°</td>
<td>TROPOMI</td>
<td>2018–2019</td>
<td>0.83</td>
<td>7.6</td>
<td>(Liu, 2021)</td>
</tr>
<tr>
<td></td>
<td>LightGBM</td>
<td>No</td>
<td>0.05°</td>
<td>TROPOMI</td>
<td>2018–2020</td>
<td>0.83</td>
<td>6.6</td>
<td>(Wang et al., 2021)</td>
</tr>
<tr>
<td></td>
<td>SWDF</td>
<td>No</td>
<td>0.01°</td>
<td>TROPOMI</td>
<td>2019–2020</td>
<td>0.93</td>
<td>4.9</td>
<td>(Wei et al., 2022b)</td>
</tr>
<tr>
<td></td>
<td>STET</td>
<td>No</td>
<td>0.1°</td>
<td>Big data</td>
<td>2013–2020</td>
<td>0.84</td>
<td>8.0</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SO₂</td>
<td>RF</td>
<td>No</td>
<td>0.25°</td>
<td>Emissions</td>
<td>2013–2014</td>
<td>0.64</td>
<td>17.1</td>
<td>(Li et al., 2020)</td>
</tr>
<tr>
<td></td>
<td>STET</td>
<td>No</td>
<td>0.1</td>
<td>Big data</td>
<td>2013–2020</td>
<td>0.84</td>
<td>10.1</td>
<td>This study</td>
</tr>
<tr>
<td>CO</td>
<td>RF–STK</td>
<td>Yes</td>
<td>0.1</td>
<td>MOPITT</td>
<td>2013–2016</td>
<td>0.51</td>
<td>0.54</td>
<td>(Liu et al., 2019)</td>
</tr>
<tr>
<td></td>
<td>LightGBM</td>
<td>No</td>
<td>0.07°</td>
<td>TROPOMI</td>
<td>2018–2020</td>
<td>0.71</td>
<td>0.26</td>
<td>(Wang et al., 2021)</td>
</tr>
<tr>
<td></td>
<td>STET</td>
<td>No</td>
<td>0.1°</td>
<td>Big data</td>
<td>2013–2020</td>
<td>0.80</td>
<td>0.29</td>
<td>This study</td>
</tr>
</tbody>
</table>

KCS: kriging-calibrated satellite method; LightGBM: light gradient boosted model; LME: linear mixed effect model; LUR: land use regression; MOPITT: Measurements of Pollution in the Troposphere; OMI: Ozone Monitoring Instrument; RF: random forest; RF-K: random forest integrated with K-means; RF-STK: random-forest-spatiotemporal-kriging model; STET: space-time extremely randomized tree; SWDF: spatiotemporally weighted deep forest; TROPOMI: TROPOspheric Monitoring Instrument; XGBoost: extreme gradient boosting