https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Sequential spatiotemporal distribution of $PM_{2.5}$, SO_2 and Ozone in China from 2015 to 2020

Yufeng Chi [a], Yu Zhan [b], Kai Wang [c], Hong Ye [d, e, f, g, *]

a. School of Information Engineering, Sanming University, Sanming 365004, China
b. Department of Environmental Science and Engineering, Sichuan University, Chengdu 610065, China
c. China-UK Low Carbon College, Shanghai Jiaotong University, Shanghai 200000, China
d. Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China
e. Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China
f. CAS Haixi Industrial Technology Innovation Center in Beilun, Ningbo 315800, China
g. Xiamen Key Laboratory of smart management of urban environment Xiamen 361021, China
Correspondence to: Hong Ye (hye@iue.ac.cn)

Abstract: Currently, in the modeling of various atmospheric pollutants, the simulation of independent trace gases (SO_2 and O_3) is constrained by the insufficient resolution of key remote sensing products, resulting in insufficient simulation reliability. In this study, spatial sampling and parameter convolution are combined to optimize LightGBM by utilizing ground observations, remote sensing products, meteorological data, assistance data, and random ID. Through the above techniques and an sequentialsimulation of air pollutants, we produce seamless daily 1-km-resolution products of $PM_{2.5}$, SO_2 and O_3 for most parts of China from 2015 to 2020. Through random sampling, random site sampling, area-specific validation, comparisons of different models, and a cross-sectional comparison of different studies, we verified that our simulations of the spatial distribution of multiple atmospheric pollutants are reliable and effective. The CV of the random sample yielded an R^2 of 0.88 and an RMSE of 9.91 μg/m^3 for $PM_{2.5}$, an R^2 of 0.89 and an RMSE of 4.62 μg/m^3 for SO_2, and an R^2 of 0.91 and an RMSE of 6.88 μg/m^3 for O_3. Combined with the SHapley Additive exPlanations (SHAP) approach, the roles of different parameters in the simulation process were clarified, and the positive role of parameter convolution was confirmed. Our dataset was used to assess the changes in the Air Pollution Index (API) in China before and after the outbreak of

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

COVID-19, and the results indicate that these changes were relatively small huge, suggesting that the epidemic control measures in 2020 were effective. The study demonstrates that the multipollutant datasets produced with the proposed models are of great value for long-term, large-scale, and regional-scale air pollution monitoring and prediction, as well as population health evaluation. The datasets are available at https://doi.org/10.5281/zenodo.7533813 (Chi et al. 2023a), https://doi.org/10.5281/zenodo.7547774 (Chi et al. 2023b), https://doi.org/10.5281/zenodo.7312179 (Chi et al. 2023c), https://doi.org/10.5281/zenodo.7580714 (Chi et al. 2023d), https://doi.org/10.5281/zenodo.7580720 (Chi et al. 2023e), https://doi.org/10.5281/zenodo.7580726 (Chi et al. 2023f).

Keywords: Multiple air pollutants, Machine learning model optimization, Spatial distribution products of air pollutants, SHAP

1 Introduction

The development of human society has led to large quantities of air pollutant emissions, seriously affecting human health (Dedoussi et al., 2020; Landrigan, 2017; Shen et al., 2019). In 2019, Global Disease Burden (GDB) data indicated that air pollution was the fourth leading cause of death. In 2015 alone, outdoor $PM_{2.5}$ and ozone (O_3) pollution caused 4.5 million deaths (Cohen et al., 2017). The concentrations of air pollutants such as $PM_{2.5}$, O_3, and SO_2 can be effectively obtained with observation devices at ground stations (World Health, 2021; Copat et al., 2020). However, due to the high cost, it is difficult to build high-density ground monitoring stations to monitor air pollutants. In areas without monitoring stations, the levels of gases that are imperceptible to the naked eye, such as O_3 and SO_2, may be misestimated, thus increasing the uncertainty of quantitative assessments of population exposure (Liu et al., 2020). Therefore, establishing a set of refined spatially distributed products related to near-surface air pollution could improve quantitative assessments of population exposure.

With the continuous development of remote sensing technology, satellite remote sensing can now be used to obtain the spatial distribution of atmospheric pollutants and has become an important scientific approach. The Ozone Monitoring Instrument (OMI) of the Aqua satellite, the SCIAMACHY sensor of ENVISAT, and the Tropospheric Monitoring Instrument (TROPOMI) of Sentinel-5P can directly observe and retrieve the levels of trace gases such as O_3 and SO_2 (Kang et al., 2021; Ialongo et al., 2020).

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Among them, the OMI is characterized by a long observation duration, sufficient data storage, and global coverage, providing key data for studies of near-surface trace gases (Xue et al., 2020). However, the low resolution of the OMI limits the application of OMI data in high-resolution simulations of trace gases. Due to the complex composition of $PM_{2.5}$, it is challenging to directly observe it through remote sensing, and it is usually necessary to combine parameters such as the aerosol optical depth (AOD) for indirect estimation. The AOD product produced from MODIS data combined with the multiangle implementation of atmospheric correction (MAIAC) algorithm provides high-resolution (1 km and daily) and stable data; additionally, this product is free and publicly available. In addition, the product can be used to recover relevant bidirectional reflectance distribution functions (BRDF) based on the time-series detection of multiangle surface features (Lyapustin et al., 2011). Compared with the traditional dark target (DT) and dark blue (DB) algorithms, it can more effectively identify clouds and snow, and the inversion effect is better in certain areas.

Since 2013, China has built several air pollutant monitoring stations, gradually laying the foundation for the establishment of a national-scale and fine-scale dataset of air pollutants (Li et al., 2017). At present, the main methods for simulating the spatial distribution of near-surface air pollutants can be categorized into physical and chemical models, mathematical and statistical models, and artificial intelligence methods (Chong et al., 2020). Physicochemical models were developed first and are often combined to form relatively complete analysis systems (such as combining remote sensing retrieval products, reanalysis data, and atmospheric chemical transport models) (Ivey et al., 2017). However, the corresponding products usually have a low resolution and cannot meet the needs of regional studies. Mathematical and statistical models include many spatial interpolation and linear algebra models (Zhang et al., 2018a). Although such models can simulate the spatial distribution of near-surface air pollutants at a high resolution, it is difficult to effectively simulate local abrupt changes (such as forest fires and abnormal emissions) (He and Huang, 2018). Therefore, this approach has not been broadly popularized and is difficult to apply over small spatial scales and in short time periods. Artificial intelligence methods, including machine learning and deep learning, have gradually matured, leading to improved simulations of the spatial distributions of atmospheric pollutants (Chang et al., 2020; Wei et al., 2022). Among them, the machine learning-based LightGBM model provides high cross-validation (CV) accuracy and reliability without requiring extensive computational resources (Ke et al., 2017; Zhong et al., 2021). However, when large-scale remote sensing data are used to simulate the

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

spatial distribution of near-surface atmospheric pollutants, especially trace gases such as SO_2 and O_3, in the LightGBM model, “bands” or “patches” that do not conform to natural patterns are often obtained (Figure S4) (Zhan et al., 2017b; Chi and Zhan, 2022). This phenomenon not only affects the reliability of the obtained spatial distributions of atmospheric pollutants but also hinders improvements to the spatial resolution of trace gas simulations. Therefore, models such as LightGBM still need to be further optimized.

Trace gases such as SO_2 and O_3 are affected by the resolution of key corresponding remote sensing products, resulting in serious constraints on the resolution and accuracy of near-surface spatial simulations (Wang et al., 2022). However, $PM_{2.5}$ data can be used to help optimize such simulations. Therefore, in this study, LightGBM is optimized using spatial sampling and parameter convolution to simulate the levels of atmospheric pollutants. Using ground observations, remote sensing products, meteorological parameters, random ID and sequential simulations of various air pollutants, the spatially distributed products of $PM_{2.5}$, SO_2, and O_3 are generated at a resolution of 1 km and at the daily scale in most of China (excluding some islands) from 2015 to 2020. We interpret the output of our model using the SHapley Additive exPlanations (SHAP) method. The air pollutant trends in China before and after the outbreak of COVID-19 are assessed using the Air Pollution Index (API). This paper is organized as follows: in Section 2, the dataset is described, Section 3 presents the methodology of the model, Section 4 presents the results of the model, Section 5 focuses on the model and its application, and Section 6 presents the conclusions.

2. Data sets

The data used in this study include daily ground monitoring data for $PM_{2.5}$, SO_2, and O_3 in China. Additionally, remote sensing data, meteorological data, and auxiliary data are used.

2.1 Air pollution monitoring data and meteorological data

In this study, hourly observation data from 2,108 air pollutant stations were obtained from January 1, 2,015, to December 31, 2,020. Among them, the National Environmental Monitoring Center of China operates 2,020 stations, the Hong Kong Environment Department operates 18 stations, and the Taiwan Environment Agency operates 70 sites. Figure 1 shows that the spatial distribution of the air pollutant monitoring sites is heterogeneous, with a higher density of stations along the east coast and a lower density in the western plateau region. In addition, we collected daily monitoring data from 760 meteorological stations in mainland China from January 1,

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

2,015, to December 31, 2,020, with a focus on four parameters: wind speed, humidity, air pressure, and temperature.

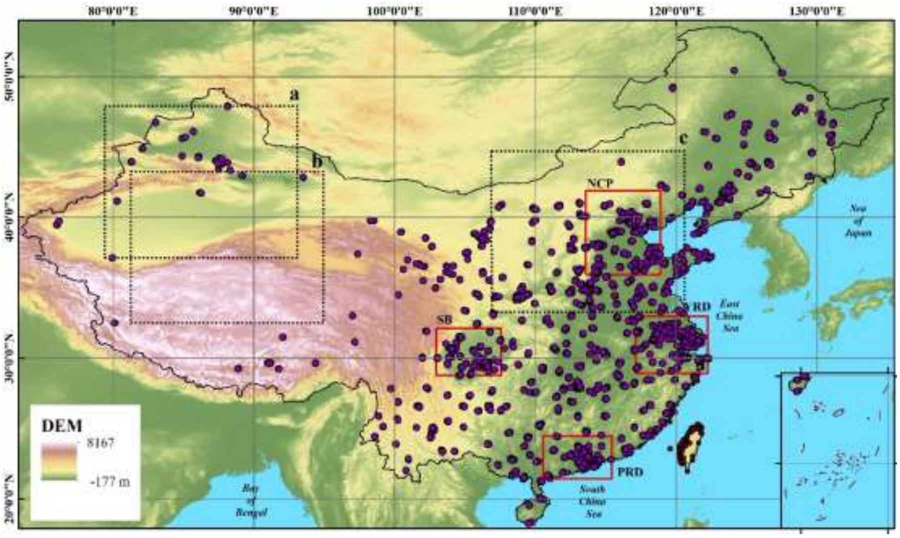


Figure 1 Map of the study area and distribution of air pollutant monitoring sites. The purple dots denote the atmospheric pollutant monitoring sites. The four red boxes represent the North China Plain (NCP), the Yangtze River Delta (YRD), the Pearl River Delta (PRD) and the Sichuan Basin (SB), areas considered in sampling CV. The three black boxes (a, b, and c) are used for visual assessment.

2.2 Remote sensing data

The remote sensing datasets used included (1) AOD datasets, (2) SO_2 and O_3 column concentration data, and (3) other datasets. (1) The MAIAC AOD and Himawari-8 AOD data sets include 470 nm AOD and 550 nm AOD. Notably, the MAIAC AOD data set (earthdata.nasa.gov) has a spatial resolution of 1 km and a temporal resolution of 1 day, and the L3 daily product of the Himawari-8 AOD data set (ftp.ptree.jaxa.jp) has a spatial resolution of 5 km. (2) The SO_2 and O_3 column concentrations are based on the L3 data for OMI SO_2 and OMI O_3, respectively, with a temporal resolution of 1 day and a spatial resolution of 0.25°. (3) Other data include NDVI, topography, population distribution, road, and land use data sets. The NDVI was calculated from MODIS data (earthdata.nasa.gov) at a temporal resolution of 16 days and a spatial resolution of 1 km. Topographic data, including elevation and slope, were extracted from SRTM data (earthdata.nasa.gov), with a spatial resolution of 90 m. Population data were obtained from LandScan (landscan.ornl.gov) at a spatial resolution of approximately 1 km. The 2018 road data were obtained from

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

OpenStreetMap (www.openstreetmap.org) in the format of an ESRI shapefile. Land use data were obtained from the Copernicus Climate Change Service (C3S) 2018, with a spatial resolution of 300 m (cds.climate.copernicus.eu).

2.3 Auxiliary data

We constructed a WGS coordinate grid covering the Chinese region (the spatial extent is shown in Figure 1) with a longitude resolution of 0.01° and a latitude resolution of 0.008°. The year parameter, day of the year (DOY) parameter, weekday/nonweekday parameter, and the independent pixel space ID parameter were considered. The data preprocessing steps are described in Data S1**.** The data description is located at Data S2.

3 Method

A general machine learning model for multiple pollutants based on random ID, spatial adoption, parameter convolution, and other methods is used to improve the consideration of multiple factors in the prediction of changes in atmospheric pollutant concentrations and optimize estimates of the spatial distributions of pollutants (Figure 2). We evaluate the model results using CV and visual qualitative analysis. LightGBM, LSTM, and RF-Ps are compared to our model to assess its performance. Finally, SHAP is used to try to interpret the output of the model.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

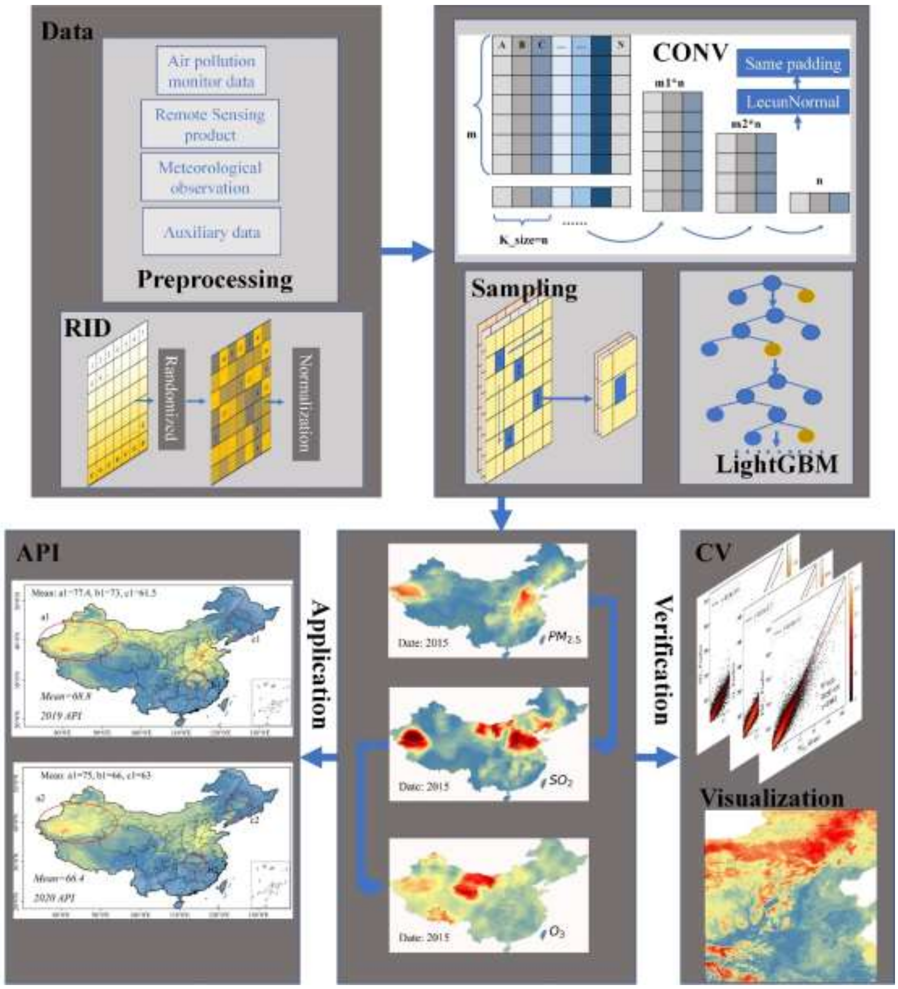


Figure 2 Technical flow chart. The diagram at the upper left shows the data collection and RID creation process. The model at the upper right includes parametric convolution, spatial sampling, and the application of LightGBM. The data are transferred to the model, and the spatial distributions of atmospheric pollutants are obtained. Then, SHAP is used to analyze the model results and generate an API for secondary analysis.

3.1 Multipollutant LightGBM model combining spatial sampling, random ID and parameter convolution

LightGBM improves upon the gradient boosting decision tree (GBDT). LightGBM mainly implements gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB). Compared with the GBDT model, LightGBM improves the calculation speed, ensures high accuracy and can better cope with large

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

amounts of data. At present, LightGBM has been applied in many fields. However, applications in atmospheric remote sensing are limited, and the potential for use in optimization is high. When developing LightGBM, we created new mechanisms for spatial sampling, parameter convolution, random ID, and the sequential simulation of multiple pollutants.

3.1.1 Spatial sampling

The spatial distribution of air pollutants is significantly affected by the locations and characteristics of monitoring sites and the surrounding environment, and many studies have considered the spatial correlations between different factors and air pollutants. We thoroughly explore the spatial information associated with remote sensing data and consider the elements near air pollutant monitoring sites. For a given pixel ($P_{(x,y)}$), the feature group of surrounding elements in a 3*3 neighborhood can be expressed as:

$$[P_{(xi,yi)}] = \{P_{(x-1,y-1)}, P_{(x-1,y)} \dots P_{(x,y+1)}, P_{(x+1,y+1)}\} \tag{1}$$

where $[P_{(xi,yi)}]$ represents an array of 8 pixel values around a given pixel ($P_{(x,y)}$).

3.1.2 Random ID

Parameter randomization is a standard model optimization method in machine learning and is widely used in various studies. The random generation of data can mitigate overfitting in the training of machine learning models and simulations involving large amounts of data. In addition, simplifying spatial feature generation can reduce the cost of model construction. Therefore, we denote the positions of all pixels with independent ID, shuffle these ID with a random algorithm, and introduce random ID (RID) into a random forest model. The specific steps are as follows.

1. Randomize the position parameters, scramble the position ID with a random algorithm, and assign a random ID to each pixel.
2. Apply a 0-1 normalization algorithm to normalize the location parameters and random location ID.

$$\begin{gathered} RID = normalization(random\ ID) \\ normalization(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \end{gathered} \tag{2}$$

where $random$ is the randomization function, x_{min} is the minimum value, and x_{max} is the maximum value.

3.1.3 Parameter convolution

The spatial distribution of air pollutants is affected by various factors, the

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

relationships among factors are complex, and the correlation coefficients among factors are low (Figure S3). In most cases, remote sensing factors do not fully reflect the many characteristics of atmospheric pollutants. To provide more features for model training, we implement random 1D convolution operations for various factors. The specific process is as follows:

1. Normalize all features.
2. Select a 1*3 convolution window.
3. Set the number of features considered for the two convolution boosting parameters, where m1=64 and m2=16.
4. Input random features into the convolution window.
5. Initialize the random convolution kernel (LecunNormal) (Klambauer et al., 2017; Lecun et al., 2012).
6. Apply the 'same padding' method to obtain a set of results.

3.1.4 Sequential simulation of multiple pollutants

$PM_{2.5}$, SO_2, and O_3 interact with each other, and there is also a solid synergistic relationship between trends in space and time. To effectively predict the spatial distribution of multiple pollutants, it is necessary to introduce different pollutants into the prediction model. We set the sequential simulation prediction order as $PM_{2.5}$>SO_2>O_3.

3.2 Other models

The LightGBM, LSTM, and RF-Ps models were used to independently simulate the spatial distributions of $PM_{2.5}$, SO_2, and O_3. Only RF-Ps included an additional parameter, namely, Ps, and the other parameters remained the same. The details of the models are given in Table 1.

Table 1 Details of the models

<table>
<tr><th>Name</th><th>Shared parameters</th><th>$PM_{2.5}$</th><th>SO_2</th><th>O_3</th><th>Special</th></tr>
<tr><td>LightGBM</td><td rowspan="3">Hum, Ws, Pr, Tem, Ele, SLOP, POP, NDVI, RL, LUCC, DOY, YEAR, WOND, PBLH, AOD_{550}, AOD_{470}, OMISO, $OMIO_3$</td><td rowspan="3">-</td><td rowspan="3">$PM_{2.5}$ Predicted</td><td rowspan="3">$PM_{2.5}$ Predicted, SO_2 Predicted</td><td>-</td></tr>
<tr><td>LSTM</td><td>-</td></tr>
<tr><td>RF-Ps</td><td>Ps</td></tr>
</table>

3.3 CV and visualization assessment

CV is divided into random CV and regular CV. Random CV is used to randomly

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

select 90% of the data for modeling and the rest for testing. This process was repeated ten times, and the average result was used. In regular CV, data from a specific time and space are used for testing, and the rest are used for training. The CV in this study were evaluated using the coefficient of determination (R^2) and root mean square error (RMSE).

Combined with atmospheric convection and regional transport theories, we qualitatively determined whether there were significant anomalies (patches and bands) in the visualization results.

3.4 Model explanation

SHapley Additive exPlanations (SHAP) is a game theory approach for calculating the importance of features in a model by comparing model estimates with and without features (Lundberg et al., 2020). A variety of parameter measurement methods can be used, and we selected the bee swarm approach to calculate the influence of each input parameter and each feature on the output (Lundberg et al., 2018). The main parameters that affect the model are identified, and the effect of each parameter on the simulation results is constrained (Zhong et al., 2021).

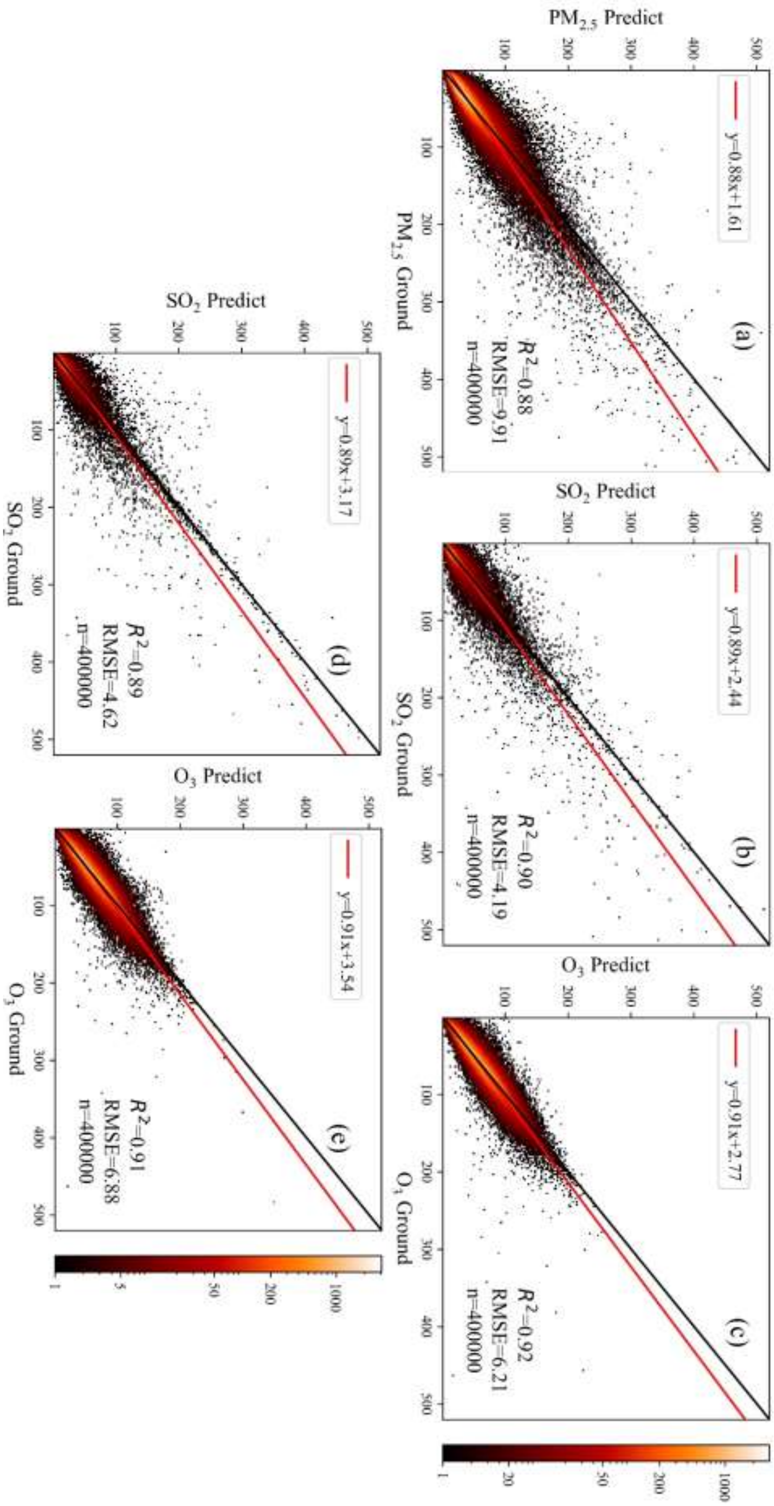
4 Results and analysis

4.1 CV results

4.1.1 Total random sampling CV

The sequential training and verification process of the models for multiple air pollutants includes training and verification using ground observation data and secondary training and verification using simulated data. Therefore, we illustrate the CV for these two steps in Figure 3.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023



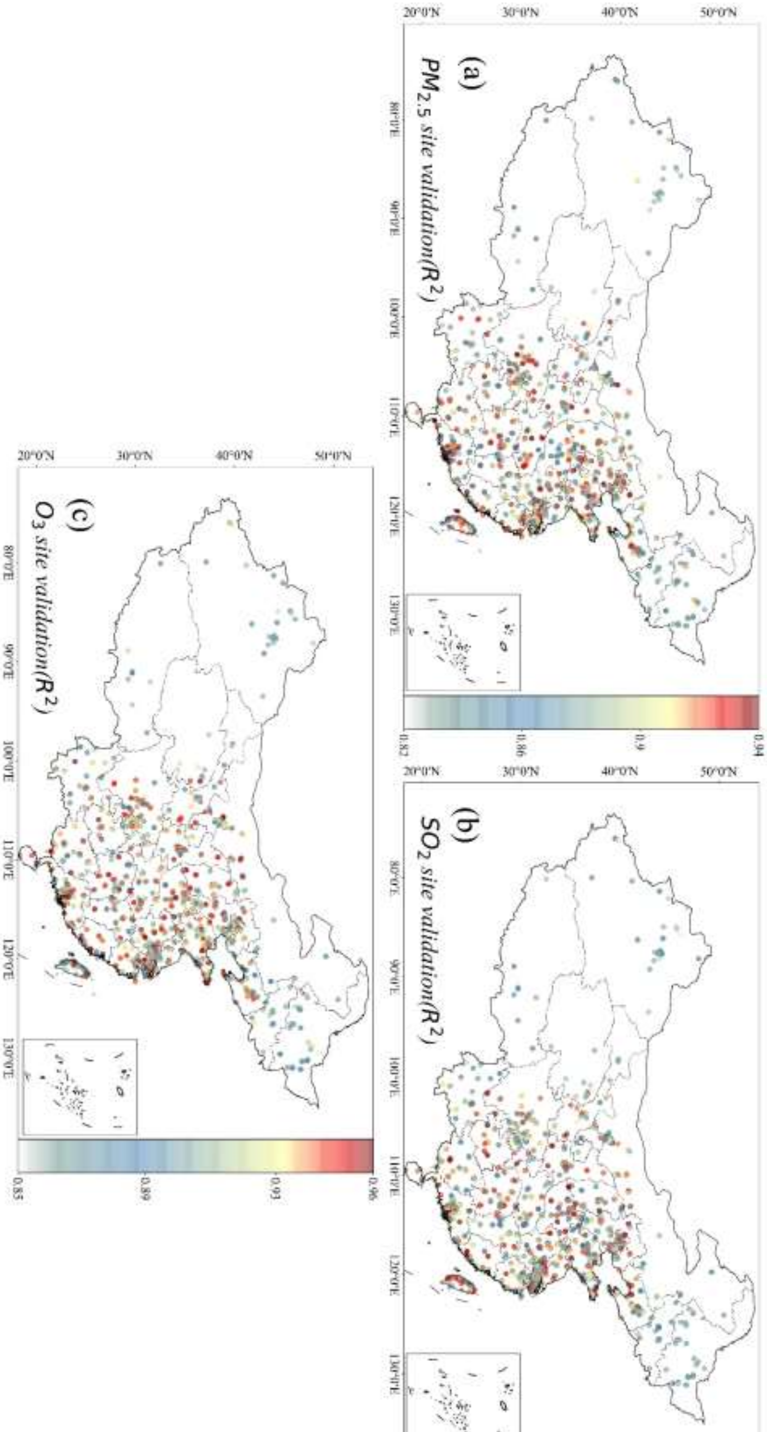
274

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Figure 3. Model construction results considering various air pollutants and CVs of the spatial distributions of pollutants. (a) CV of $PM_{2.5}$ in the model. (b) CV of SO_2 model trained with $PM_{2.5}$ ground observation. (c) CV of O_3 model trained with SO_2 ground observation. (d) CV of SO_2 model trained with PM2.5 simulation. (e) CV of O_3 model trained with SO_2 simulation. In the figure, n represents the number of samples, and the color bar on the right represents the density of the samples. The black line represents the 1:1 reference. The red line represents the results of sample fitting.

The estimation model of SO_2 uses $PM_{2.5}$ ground observation data, and the O_3 model uses $PM_{2.5}$ and SO_2 ground observation data. However, the lack of complete spatial information of air pollutants, this process cannot achieve further spatial modeling of multiple air pollutant products. Therefore, in the spatial distribution model, the predicted spatial air pollutants are used as the model inputs. For example, the estimation model of SO_2 uses the simulated spatial distribution of $PM_{2.5}$. Figure 3 shows that as the number of parameters increases, the R^2 of $PM_{2.5}$, SO_2, and O_3 increase sequentially. In addition, the estimates of the models based on simulation results are slightly lower than the site observations by approximately 1% (SO_2 and O_3).

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023



https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Figure 4 Random site sampling verification results for $PM_{2.5}$, SO_2 and O_3. The dots represent the spatial locations of the monitoring stations, and the colored column denotes the R^2.

We randomly sampled one-tenth of the site data for CV (Figure 4). The R^2 of $PM_{2.5}$, SO_2, and O_3 varied between 0.82-0.94, 0.84-0.95, and 0.85-0.96, respectively. In addition, R^2 were higher in regions with a dense station distribution and lower in regions with a sparse station distribution (such as western China).

4.1.2 Regular sampling CV

The North China Plain (113.6°E-118.8°E, 36°N-41.9°N), Yangtze River Delta (117°E -122.2°E, 29°EN-32.9°N), Pearl River Delta (110.4° E-115.3°E, 21.5°N, 24.6°N), and Sichuan Basin (102.9°E-107.5°E, 28.8°N -32.2°N) were selected for CV analysis. The CV verifications of the $PM_{2.5}$, SO_2, and O_3 simulation models in different regions were performed separately (Figure 5).

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Open Access Earth System Science Data Discussions

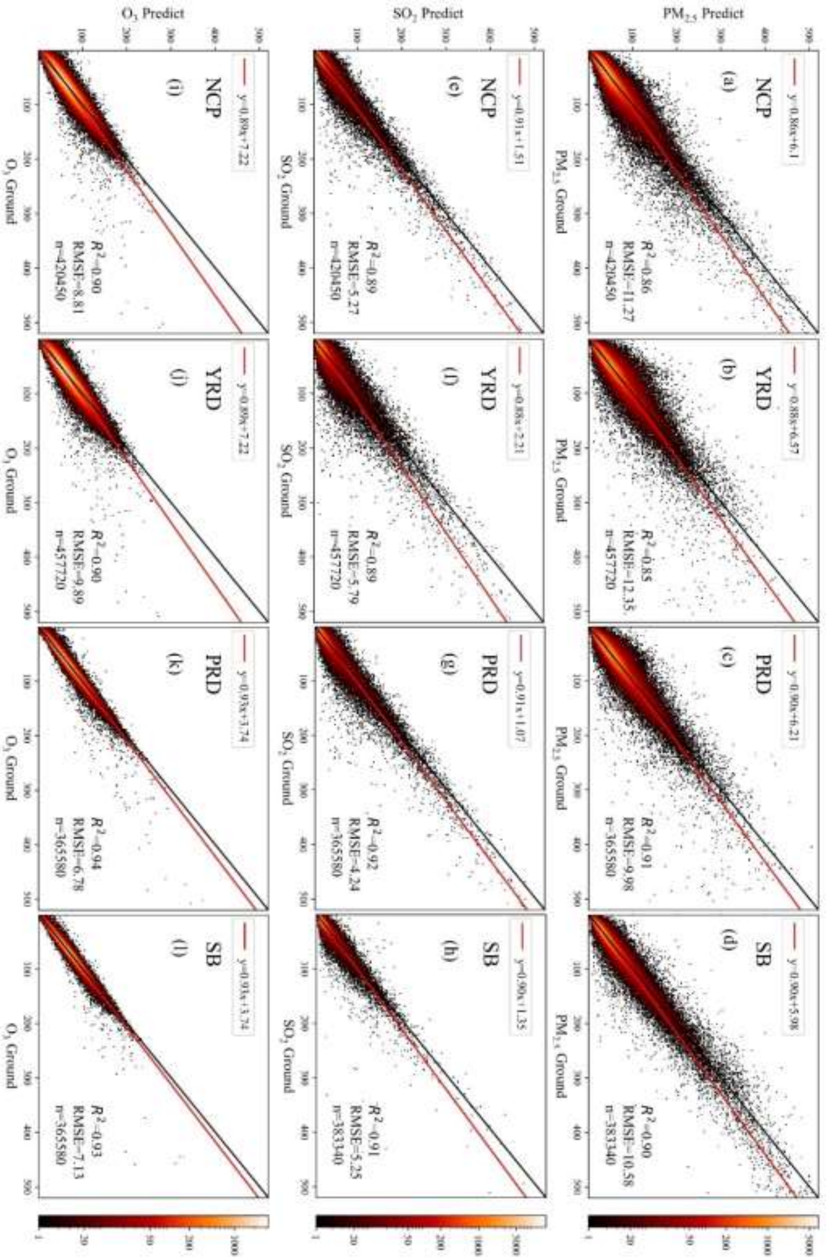


Figure 5. CV of $PM_{2.5}$, SO_2, and O_3 in different regions. The simulation mode refers to using the simulation data as an input. a to d show the results of the four-region $PM_{2.5}$ CV. e to h show the results of the SO_2 CV in the four regions. i to l show the results of the O_3 CV in the four regions. NCP, YRD, PRD, and SB denote the North China Plain, Yangtze River Delta, Pearl River Delta, and Sichuan Basin, respectively.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Figure 5 shows that satisfactory RMSE and R^2 are obtained for the sampling results in the four regions. Notably, the R^2 for $PM_{2.5}$, SO_2, and O_3 sampling in the NCP and YRD regions are lower than those in the PRD and SB, and the RMSE are higher. The reason for these differences may be related to the amounts of training data and validation data used. However, the results verify the stability of the proposed model in regional validation (regular spatial sampling).

Next, the data from each month and each year were sampled as validation samples, and the model was retrained. The corresponding CV statistics are shown in Figure 6.

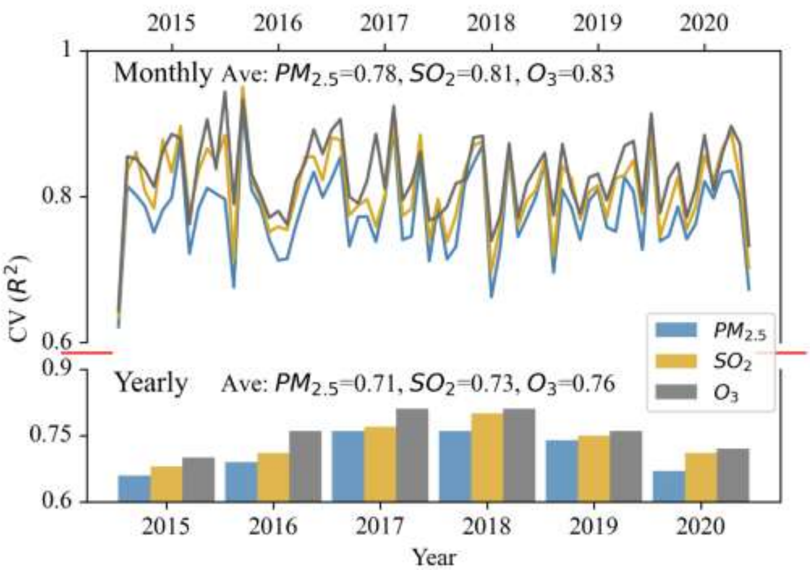


Figure 6. Annual and monthly CV of samples from 2015-2020. The upper part of the figure shows the mean of the resulting curve and CV for monthly sampling, and the lower part of the figures illustrates the bar plots and means for annual sampling. The three colors of the curves and columns denote $PM_{2.5}$, SO_2, and O_3.

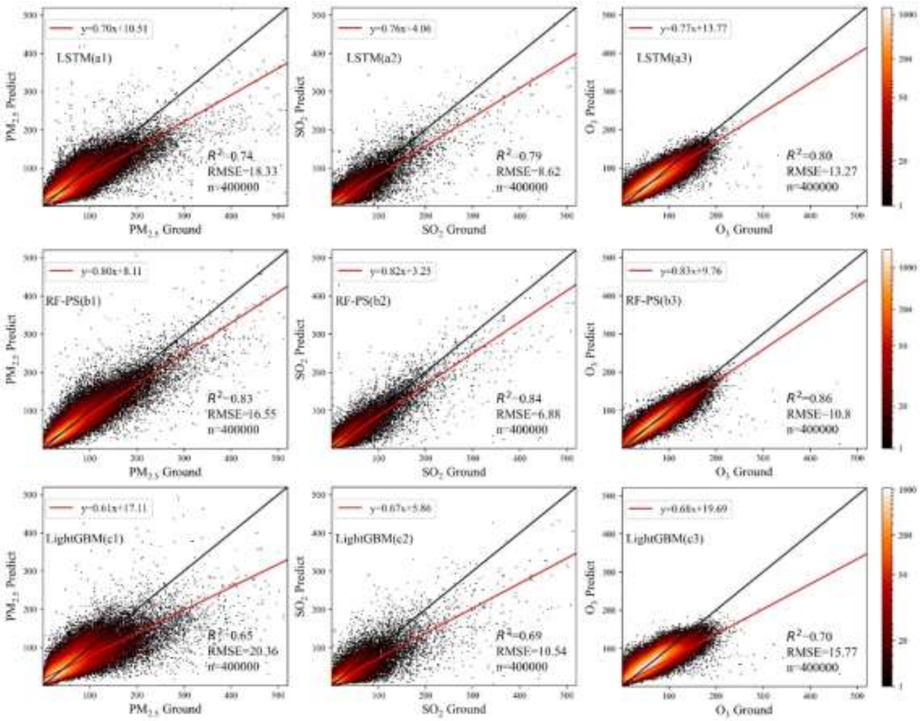
In Figure 6, the R^2 of the monthly sampling for $PM_{2.5}$, SO_2, and O_3 is not as high as that for random sampling but is similar (0.78-0.83). The R^2 for $PM_{2.5}$, SO_2, and O_3 based on monthly sampling are all higher than those for annual sampling (0.71-0.76); this result is related to the number of samples considered for training and validation. Regardless of whether the three pollutants were sampled monthly or annually, the average R^2 displayed the following order: $PM_{2.5}$<SO_2<O_3. Compared to random and regular spatial validation, regular temporal sampling validation was associated with lower R^2, especially for CV at the annual scale. However, the model still displayed strong stability.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

4.1.3 CV of LSTM, RF-Ps, and LightGBM

Figure 7 shows the CV of random sampling for the LSTM, RF-Ps, and LightGBM models.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023



https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Figure 7 CV of the LSTM, RF-Ps, and LightGBM models. LSTM(a1)-LSTM(a3) illustrate the CV of $PM_{2.5}$, SO_2, and O_3 simulations using the LSTM model, RF-Ps(b1)-RF-Ps(b3) show the CV of $PM_{2.5}$, SO_2, and O_3 simulations using the RF-Ps model, and LightGBM(c1)-LightGBM(c3) illustrate the CV of $PM_{2.5}$, SO_2, and O_3 simulations using the LightGBM model.

In Figure 7, the CV of the LSTM and RF-Ps models are similar to those of the proposed model for $PM_{2.5}$, SO_2 and O_3, with $R^2(PM_{2.5}) < R^2(SO_2) < R^2(O_3)$. This result suggests that air pollutant output data can be input into different models to improve the predictions of other pollutants. However, the R^2 and RMSE obtained for the LSTM and RF-Ps models are quite different from those of our model. Among the three models, the best CV are obtained for RF-Ps. However, our model still yields the highest R^2 and RMSE. Notably, the R^2 value of the proposed model is approximately 5% higher than that of the RF-Ps model. Additionally, the RMSEs of the proposed model are 2 μg/m^3, 2.3 μg/m^3, and 4 μg/m^3 lower than those of the RF-Ps model for $PM_{2.5}$, SO_2, and O_3, respectively. The LightGBM model performs poorly based on both the R^2 and RMSE, possibly due to the lack of auxiliary parameters and optimization. Comparatively, our model and the RF-Ps model use more auxiliary parameters than LightGBM, indicating that artificial auxiliary parameters enhance model training. Compared with the RF-Ps model, our model mainly improves the parameter convolution process and uses parameter convolution to further explore the relationships among features and parameters. Although the LSTM model does not perform as well as our model based on various verification parameters, it displays excellent development potential.

In addition, we performed CV assessments of the random sampling approach after adding RID, Ps, and RID+Ps parameters to LightGBM (Figure S5). The results indicated that the RID increased the performance of LightGBM more so than did Ps and RID+Ps, suggesting that the RID are the most stable input parameters.

We measured the time required to run the 4 models, as shown in Table 2 (for the $PM_{2.5}$ case).

Table 2 Time efficiency of the four models

Name	Time ratio	R^2 ($PM_{2.5}$)	GPU
LightGBM	1	0.65	available
RF-Ps	12.56	0.83	unavailable
LSTM	7.5	0.74	available
Ours	1.95	0.88	available

In terms of efficiency, LightGBM runs the fastest, followed by our model, with

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

the LSTM and RF-Ps models required much more time to run. Among them, LightGBM, the LSTM model and our model all support GPU computing. However, RF-Ps is not yet supported on GPUs (Kim et al., 2021). In addition, we selected 16 models from the relevant literature to compare with our model based on CV, RMSE, and spatial resolution results, and the findings are presented in Table 3.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Table 3. Comparison of multiple models in the simulation of different air pollutants

$PM_{2.5}$				SO_2				O_3			
Name	R^2	RMSE	Resolution	Name	R^2	RMSE	Resolution	Name	R^2	RMSE	Resolution
(You et al., 2016)	0.79	18.6	3 km	(Li et al., 2019)	0.62	10.36	0.25°	(Wang et al. 2022	0.84		10 km
(Xiao et al., 2018)	0.79	21		(Zhang et al., 2019)	0.64	19.5	0.1°	(Watson et al., 2019)	0.67		
(Zhang et al., 2018b)	0.85	12.4	10 km	(Zhang et al., 2021)	0.74	10.49	0.25°	(Han et al., 2022)	0.65		
(Chen et al., 2019)	0.86	14.98	3 km	(Devi et al. 202	0.74	12.6		(Zhan et al., 2018)	0.69	26	0.1°
(Zhan et al., 2017a)	0.76	23	1 km					(Silibello et al., 2021)	0.8		1 km
(Wei et al., 2019)	0.85	15.57	1 km					(Liu et al., 2020)	0.78	21	0.1°
RF-Ps	0.83	16.55	1 km	RF-Ps	0.84	6.88	1 km	RF-Ps	0.86	10.8	1 km
LSTM	0.74	18.33	1 km	LSTM	0.79	8.62	1 km	LSTM	0.8	13.27	1 km
LightGBM	0.65	20.36	1 km	LightGBM	0.69	10.54	1 km	LightGBM	0.7	15.77	1 km
Ours	0.88	9.91	1 km	Ours	0.89	4.62	1 km	Ours	0.91	6.88	1 km

In Table 3, compared with recent machine learning models, the proposed model yields better results for $PM_{2.5}$, SO_2, and O_3

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

4.2 Visual comparison of the spatial distribution of air pollutants

We randomly sampled the spatial distributions of $PM_{2.5}$, SO_2, and O_3 on January 26, 2,015, and performed corresponding simulations with the LSTM, RF-Ps, and LightGBM models.

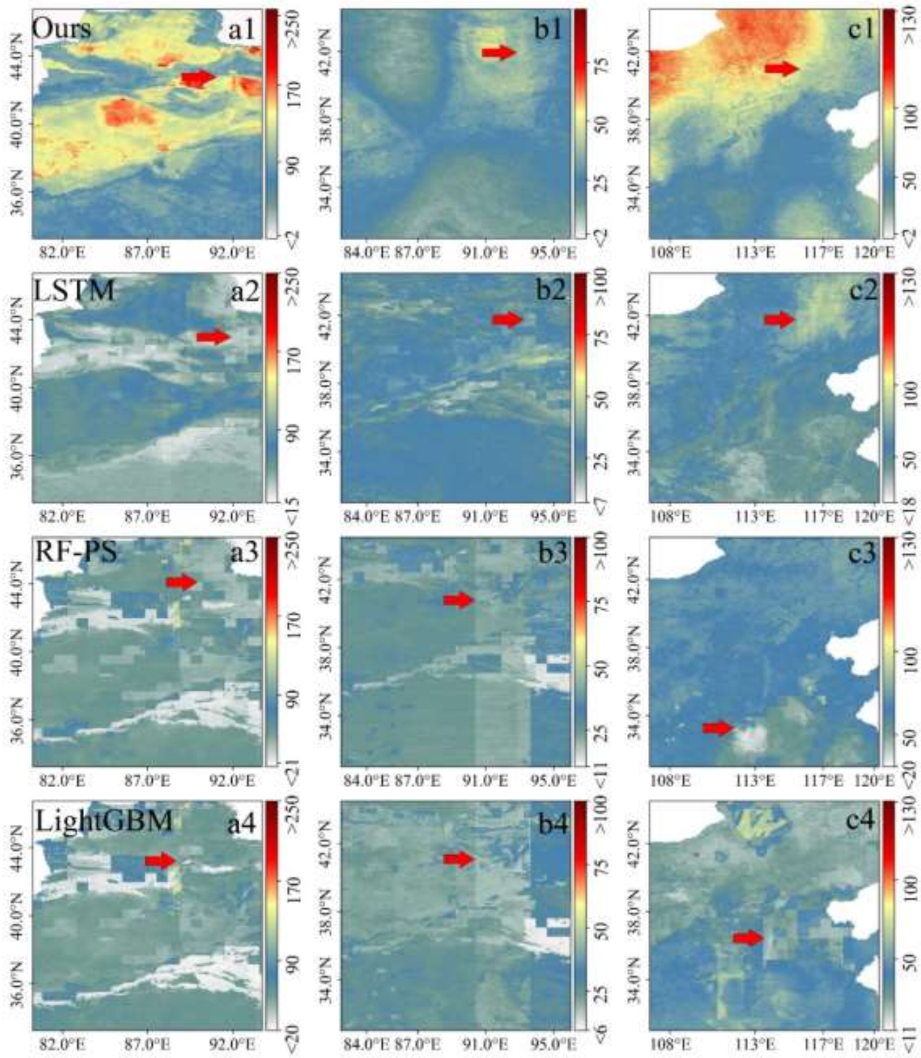


Figure 8. Local comparison of different methods. a1, a2, a3, and a4 illustrate the $PM_{2.5}$ results of our model, LSTM, RF-Ps, and LightGBM, respectively. b1, b2, b3, and b4 illustrate the SO_2 results of our model, LSTM, RF-Ps, and LightGBM, respectively. c1, c2, c3, and c4 illustrate the O_3 results our model, LSTM, RF-Ps, and LightGBM, respectively. The red arrows indicate whether there is an abnormal spatial distribution in the local area. The red bars represent

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

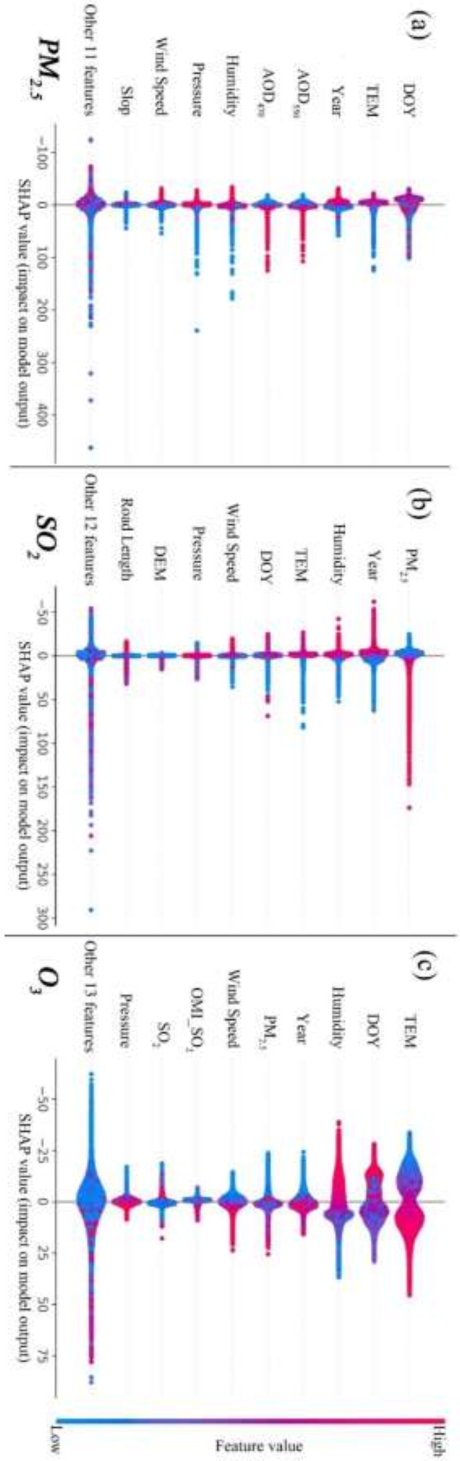
atmospheric pollutant concentrations.

The red arrows in Figure 8 indicate the anomalies observed in the simulation of pollutant distributions in local areas and bands. For the results in a1, b1 and c1, which were obtained with our model, few anomalies are present. Additionally, the visualization effect of LSTM is better than that of RF-Ps and LightGBM.

4.3 SHAP results

Figure 9 shows results of the SHAP approach with the bee swarm method, which was used to assess the impact of each sample and parameter on the model results. Moreover, SHAP was used to analyze the influence of parameter convolution on the model results (Figure 10).

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023



https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Figure 9. SHAP bee swarm results. a, b, and c show the SHAP results for $PM_{2.5}$, SO_2, and O_3, respectively. The color bar on the right represents the relative magnitude of the variable value, and the abscissa represents the SHAP value.

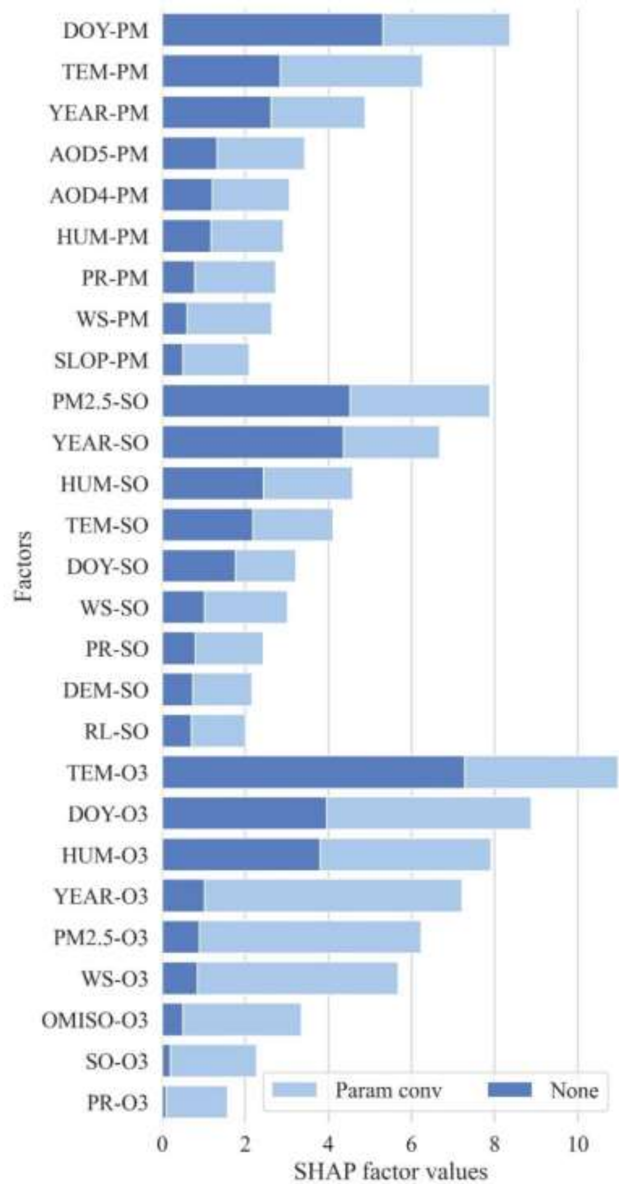


Figure 10. Comparison of the SHAP values with and without applying parameter convolution. PM represents the main parameters used to simulate $PM_{2.5}$, SO represents the main parameters used to simulate SO_2, O_3 represents the main parameters used to simulate O_3, Param conv represents the use of parameter convolution, and None indicates the absence of parameter convolution.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Figure 9 shows the SHAP summary for the proposed model, and the ranking of features from top to bottom reflects the importance of each feature in the model. The results show that different variables have different effects on the simulation of $PM_{2.5}$, SO_2, and O_3. We note that in our model, DOY and Year are crucial when constructing air pollutant models. Notably, DOY air pollutant simulations are comparatively random, and Year is negatively correlated with $PM_{2.5}$ and SO_2 and positively correlated with O_3. The influence of the Year parameter on the model corresponds to the gradual improvement of the air pollution status in China in recent years. Meteorological parameters are also critical and relatively strongly related to the physical and chemical relationships among and spatial distribution of atmospheric pollutants. For example, the lower (higher) the temperature is, the higher (lower) the $PM_{2.5}$ level; the lower (higher) the wind speed is, the higher (lower) the SO_2 level; and the lower (lower) the humidity is, the higher (lower) the O_3 level. In addition, pollutant parameters significantly affect the simulation of $PM_{2.5}$, SO_2, and O_3. For example, AOD has a significant positive effect on the simulation of $PM_{2.5}$, and $PM_{2.5}$ displays a similar effect in SO_2 simulations. Moreover, $PM_{2.5}$, SO_2, and OMISO simulation results all influence O_3 prediction.

In Figure 10, the SHAP value is the mean absolute value of the SHAP value of each sample, and the larger the value is, the stronger the contribution of the parameter to estimates of the concentrations of atmospheric pollutants. Notably, the convolution parameter significantly contributes to improvements in the predictions of atmospheric pollutants.

4.4 Long-term spatial distribution characteristics of various air pollutants

Figure 11 shows the average annual distributions of $PM_{2.5}$, SO_2, and O_3 in China from 2015 to 2020 simulated with the proposed method.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

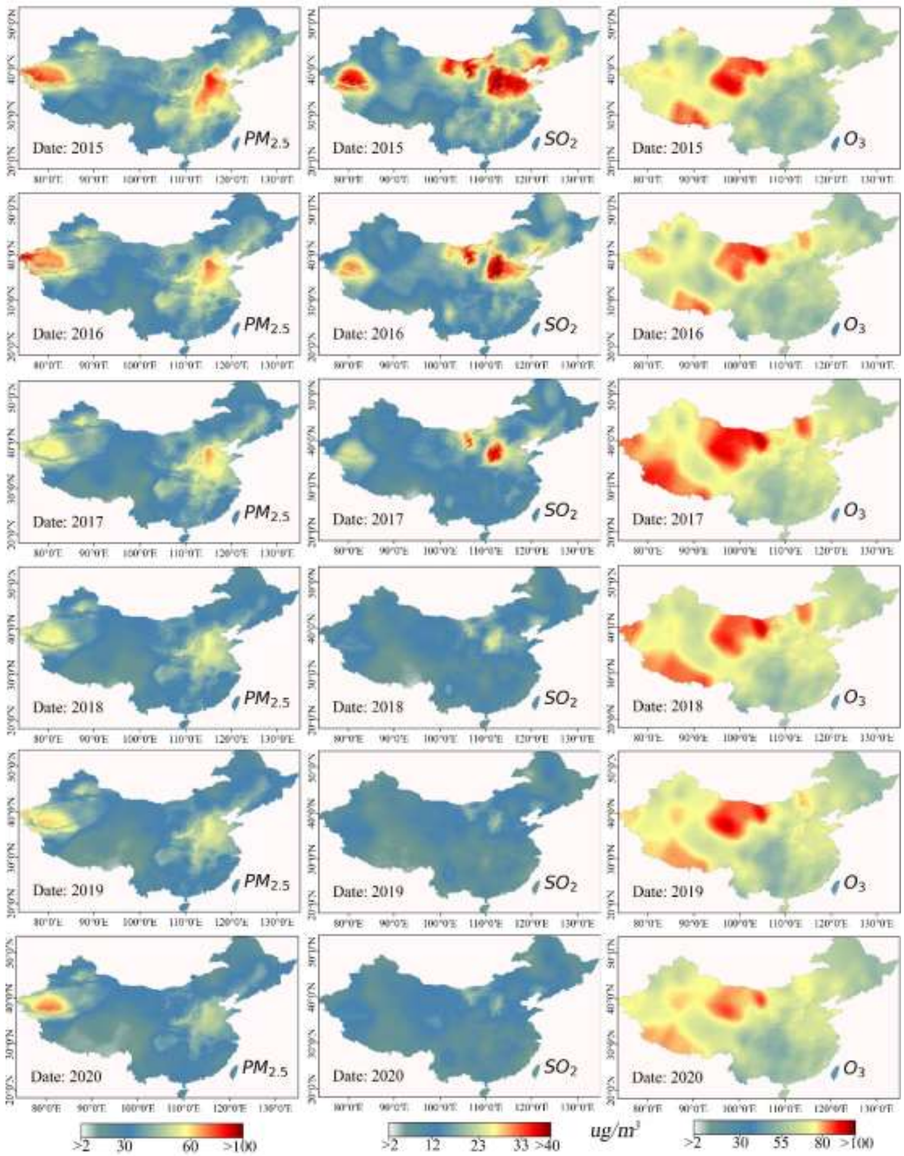


Figure 11. Maps of the annual average spatial distributions of $PM_{2.5}$, SO_2, and O_3 in China from 2015 to 2020. a1-a6 show the annual average $PM_{2.5}$ values from 2015-2020. b1-b6 show the annual average SO_2 values from 2015-2020. c1-c6 illustrate the annual average O_3 values from 2015-2020. The bar at the bottom gives the concentrations of pollutants in the study area.

The high-risk areas of $PM_{2.5}$ and SO_2 are mainly located in the northern and northwestern parts of China. Although ozone is also high in these two regions, there are two high-value areas in northern and northwestern China and on the Qinghai-Tibet

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Plateau. The findings of Gao et al. (Gao et al., 2020; Zhong et al., 2021; Zhang et al., 2019), $PM_{2.5}$, SO_2 and O_3 further confirm the reliability of our results.

4.6 Impact of COVID-19 on air pollution in China in 2019 and 2020

Changes in air pollution before and after the COVID-19 pandemic can be effectively assessed using the API. Based on the calculation method reported in the National Environmental Protection Standard of the People's Republic of China - Ambient Air Quality Index (AQI), we calculated the daily API values of $PM_{2.5}$, SO_2, and O_3 in 2019 and 2020. Figure 12 shows the average annual spatial distribution of the API in 2019 and 2020. If the API exceeds 100, it means that the day has exceeded the secondary standard of ambient air pollution concentration limit. Figure 13 shows the number of days on which the API exceeded 100.

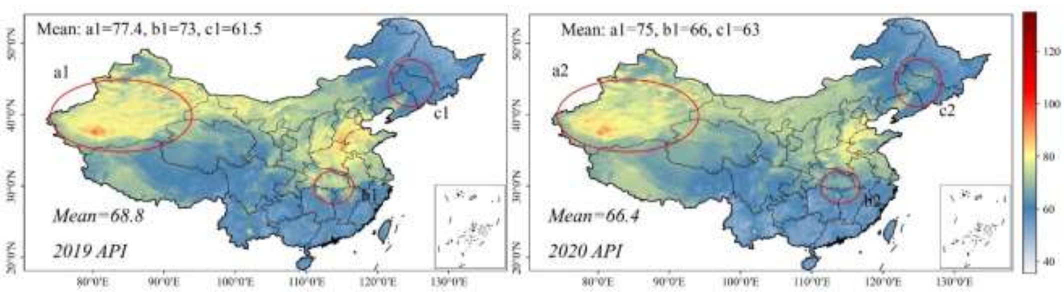


Figure 12. Spatial distribution of API in China in 2019 and 2020. a shows the results for the Xinjiang region of China, with an API of 77.4 in 2019 and 75 in 2020. b shows the results for Hubei, China. Wuhan was on lockdown for the first time due to COVID-19 from January 23 to April 8, 2020. The API was 73 in 2019 and 66 in 2020. c shows the results for the Jilin region in Northeast China, with an API of 61.5 in 2019 and 63 in 2020. The color bar on the right shows the magnitude of the API values.

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

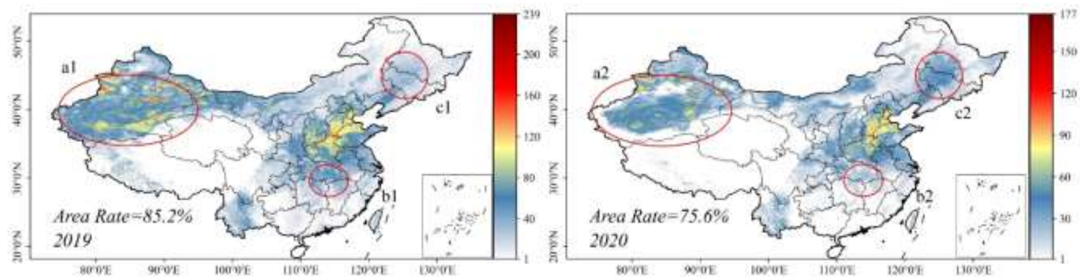


Figure 13. Spatial distribution of the number of days with API values over 100 in China in 2019 and 2020. In the white regions, the API was less than 100 each day during the study period. The maximum number of days with API values exceeding 100 in China was 239 in 2019 and 177 in 2020.

The results in Figure 12 and 13 are consistent with the trend of decreasing concentrations of major air pollutants in China. The API in China in 2019 and 2020 displayed a downward trend, decreasing from 68.8 in 2019 to 66.4 in 2020. The percentage of areas with API values greater than 100 decreased from 85.2% in 2019 to 75.6% in 2020. The number of days with an API over 100 also decreased from 239 to 177 days. The influence of the main pollutant $PM_{2.5}$ gradually decreased, and the range of influence of O_3 increased. In addition, the API in central China declined in 2020, the API in the northwest nonsignificantly decreased, and the API in the northeast increased (Wen et al., 2020).

In the obtained histogram and the API results (Figure S6), both the maximum value and the average value of the API decreased from 2019 to 2020, but the API values generally remained high. Since 2015, $PM_{2.5}$ and SO_2 have displayed significant downward trends, but the downward trend of O_3 is not apparent (Figure 9 and Figure 10). As shown in Figures 11-13, the epidemic in 2020 had a significant impact on air pollution in local areas (such as Wuhan and Hubei). However, the impact on the entire region of China is not particularly obvious. Due to the closure of Wuhan and other effective control measures in the early stage of the epidemic, the restriction of human activities significantly reduced air pollution in some areas in 2020. However, these measures in specific cities did not influence trends in the rest of China. In the second half of 2020, with the global spread of the epidemic, the industrial chains in other parts of the world were severely impacted, which in turn led to an increase in the industrial production capacity in areas of China not affected by the epidemic, thus increasing the emission of air pollutants to a certain extent. Local lockdowns associated with epidemic led to the return of urban workers to their hometowns, increased straw burning (remote

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

sensing observations suggest that the number of fires in 2020 increased by 20% over the number in 2019) (Meeprc, 2020, 2021), increased domestic heating and other phenomena that have exacerbated air pollution in Northeast China and other regions. Still, under the governance of policies such as the "Battle of Blue Sky and White Clouds", the air pollution conditions in China have generally improved since 2020.

5 Discussion

In-depth explorations of the spatial and temporal distributions of air pollutants will help enhance the understanding of the relationship among regional ecological security, population health, and air pollutants. Machine learning models can be used to effectively predict the spatial distributions of atmospheric pollutants. In this study, random ID, spatial sampling, parameter convolution, and the sequential simulation of various air pollutants are used to further optimize the accuracy of the proposed machine learning model to simulate the spatial distributions of air pollutants.

5.1 Model overview

This study introduces a variety of optimization rules based on LightGBM, ground air pollutant observations, and remote sensing, meteorological, and auxiliary data. Following sequential model training, gap-free $PM_{2.5}$, SO_2, and O_3 products were obtained at a 1 km daily resolution near the ground in China. Good results were achieved for $PM_{2.5}$ (R^2=0.88, RMSE=9.91 μg/m^3), SO_2 (R^2=0.89, RMSE=4.62 μg/m^3), and O_3 (R^2=0.91, RMSE=6.88 μg/m^3). Additionally, the optimization processes applied did not seriously hinder the efficiency of the model.

5.2 The efficacy of the model

Simulations of the spatial distributions of air pollutants require remote sensing data. The accuracy and resolution of remote sensing data largely influence the CV and visualization of atmospheric pollutant results (Colmer et al., 2020). Due to the limited variety and quantity of remote sensing products, it is important to construct new parameters and effectively use known parameters. Notably, the use of the Ps parameter can improve the CV of models, such as RF-Ps and LightGBM+ Ps. However, the Ps parameter does not enhance the visualization of results. Alternatively, RID can enhance the CV process and visualization of results, mainly because each pixel is associated with an independent ID. The independent ID can be used to optimize the impact of low-resolution remote sensing products on the model and then mitigate the patch or banding phenomenon. Spatial sampling and parameter convolution are two ways to effectively utilize existing parameters. Spatial sampling can provide valuable spatial domain information for each parameter, and parameter convolution can combine features associated with different parameters. The results show that under the premise of enhancing CV, the stability and generalization ability of the model can be further improved with RID and random sampling, and patch and banding phenomena are avoided.

Based on the SHAP approach, the influence of different parameters on a model can be clearly expressed, and the positive or negative effect of a given sample or parameter can be visualized. Many physical variables (such as TEM for O_3, $PM_{2.5}$ for SO_2, and AOD for $PM_{2.5}$) have significant effects on air pollutant levels (positive or negative), and nonphysical variables such as DOY exhibit certain positive or negative correlations with air pollutant levels. Although the impact on air pollutants is significant in most cases, the correlation is not consistently positive or negative. This is mainly because nonphysical variables are related to anthropogenic activities and are much more random than physical variables. These factors should be considered in further assessments of air pollution based on machine learning simulations.

In addition, the SHAP approach was used to assess the role of parameter convolution in the proposed model. Parameter convolution can be employed to efficiently use existing data and improve the modeling of atmospheric pollutants by considering different parameters.

The selection of parameters in machine learning models should be performed with caution, and blind selection may degrade the overall performance of the model (Figure S4). There are obvious correlations among air pollutants, and understanding these relations can enhance the construction and application of air pollutant models. Specifically, one way to improve the simulation of trace gases near the surface is to fully utilize $PM_{2.5}$ simulation results. In this study, with the addition of atmospheric pollutant parameters, the CV of the SO_2 and O_3 models were enhanced. However, the repeated use of simulated atmospheric pollutants increases uncertainty to some extent. Therefore, the proposed model was only used to simulate three air pollutants. In the future, we will conduct in-depth research to quantify and resolve the uncertainties in atmospheric pollutant simulations and then simulate additional major atmospheric pollutants.

In addition to changes involving the data used, a more powerful deep learning model should be developed in the future. However, first, the fitting effect of LSTM must be improved in the context of this study, although the CV results were better than those of LightGBM. Shwartz et al. and Grinsztajn et al. (Grinsztajn et al., 2022;

Shwartz-Ziv and Armon, 2022) noted that in the processing of tabular data, most models are inferior to machine learning models, which is one of the reasons why the performance of the LSTM model is not ideal in this study. However, simulations of the spatial distributions of atmospheric pollutants are limited to tabular data supported by remote sensing products and other graphical data. We have shown that spatial sampling and parametric convolution are effective steps when using these types of data, and both of these steps are closely related to convolutional methods in deep learning. Moreover, the characteristics of input data should be considered when new parameters are selected, and blind selection should be avoided. In the future, we will combine time series and graphical neural networks to further explore the spatial distribution of air pollution.

5.3. Limitations and prospects

1) The TROPOMI mounted on the Sentinel-5P satellite can obtain SO_2 and O_3 data at a higher spatial resolution than that provided by the OMI. Unfortunately, these data were last provided in 2018. We believe that using more recent data in subsequent research as they become available will further improve the accuracy of simulations of atmospheric pollutants such as SO_2 and O_3.

2) The limited accuracy of regular CV at the annual scale may limit predictions of the spatial distributions of air pollutants in the past or the future. Therefore, further improving the accuracy of annual and long-term atmospheric pollutant simulations will be a focus of our research.

3) The critical indicator used in $PM_{2.5}$ simulations is AOD, and the temporal resolution of AOD data obtained with geostationary satellites is less than one hour. Therefore, the spatial distribution of $PM_{2.5}$ simulations can be obtained at the hourly scale. However, the OMI or TROPOMI cannot achieve this resolution. The sequential simulation of atmospheric pollutants can provide similar inputs to obtain predictions of the levels of other atmospheric pollutants. Therefore, it is important to reduce the uncertainty associated with the sequential simulation of air pollutants, improve the spatial distributions of major air pollutants such as PM_{10}, NO_2, and CO, and effectively estimate the spatial distribution of the AQI. In the future, we will publish our products and codes at (https://github.com/pingyinforbidden/china_air_pollutions).

6 Data availability:

Spatial distribution of various air pollutants in China at 1 km in this manuscript can be accessed at repository under data dois:

Table 4 Data DOIs

Name	DOI	Citation

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

Earth System Science Data Open Access Discussions

$PM_{2.5}$	https://doi.org/10.5281/zenodo.7533813	(Chi et al. 2023a)
	https://doi.org/10.5281/zenodo.7547774	(Chi et al. 2023b)
SO_2	https://doi.org/10.5281/zenodo.7312179	(Chi et al. 2023c)
	https://doi.org/10.5281/zenodo.7580714	(Chi et al. 2023d)
O_3	https://doi.org/10.5281/zenodo.7580720	(Chi et al. 2023e)
	https://doi.org/10.5281/zenodo.7580726	(Chi et al. 2023f)

7 Conclusion

We introduced RID based on multisource heterogeneous data. The spatial sampling method and parameter convolution function were applied to improve the performance of LightGBM. Using the above approach combined with sequentialsimulation, daily gap-free $PM_{2.5}$, SO_2, and O_3 products were obtained with a spatial resolution of 1 km in most areas of China from 2015 to 2020. Based on random sampling CV for the proposed model, we obtained an R^2 of 0.88 and an RMSE of 9.91 μg/m^3 for $PM_{2.5}$, an R^2 of 0.89 and an RMSE of 4.62 μg/m^3 for SO_2, and an R^2 of 0.91 and an RMSE of 6.88 μg/m^3 for O_3. In addition, we demonstrated the stability and excellent generalization ability of our model by utilizing random sampling site validation, rule validation, and side-by-side comparison. We obtained 1 km of daily simulated products for $PM_{2.5}$, SO_2 and O_3. In the visualization validation, it was confirmed that our model reduced the insufficient visualization of patches and bands, even when simulating the spatial distribution of multiple pollutants in the large-scale study area. We also introduced the SHAP method to quantitively verify the optimization effect of parameter convolution in the model and assess effects of different parameters on the simulated spatial distributions of atmospheric pollutants. The results indicated that LightGBM with RID, spatial sampling, parameter convolution and sequential simulation was able to effectively and stably simulate the spatial distributions of various atmospheric pollutants. Finally, we used the simulated air pollutant data to regenerate the spatial distribution of the API and assess the corresponding trends in most regions of China in 2019 and 2020. The method proposed in this paper is of great significance for comprehensive high-resolution, large-area simulation research involving the spatial distributions of various atmospheric pollutants.

Author contributions

Y C: collected and processed the data, designed the model and wrote the manuscript. Y Z, K W and H Y revised the manuscript. All authors contributed to the study.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgments : This research was funded by National Natural Science Foundation of China (grand No.42271299), International Partnership Program of Chinese Academy of Sciences (grand No. 132c35kysb2020007), Ningbo Commonweal Science and Technology Planning Project (grand No. 2021S081) and Introduction of high-level talents in Sanming University (RD21006P, KC22020P).

References

Chang, F. J., Chang, L. C., Kang, C. C., Wang, Y. S., and Huang, A.: Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques, Sci Total Environ, 736, 139656, 10.1016/j.scitotenv.2020.139656, 2020.

Chen, Z.-Y., Zhang, T.-H., Zhang, R., Zhu, Z.-M., Yang, J., Chen, P.-Y., Ou, C.-Q., and Guo, Y.: Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China, Atmospheric Environment, 202, 180-189, https://doi.org/10.1016/j.atmosenv.2019.01.027, 2019.

Chi, Y. and Zhan, Y.: A Simple and Effective Random Forest Refit to Map the Spatial Distribution of NO2 Concentrations, 10.3390/atmos13111832, 2022.

Chong, H., Lee, S., Kim, J., Jeong, U., Li, C., Krotkov, N. A., Nowlan, C. R., Al-Saadi, J. A., Janz, S. J., Kowalewski, M. G., Ahn, M.-H., Kang, M., Joiner, J., Haffner, D. P., Hu, L., Castellanos, P., Huey, L. G., Choi, M., Song, C. H., Han, K. M., and Koo, J.-H.: High-resolution mapping of SO2 using airborne observations from the GeoTASO instrument during the KORUS-AQ field study: PCA-based vertical column retrievals, Remote Sensing of Environment, 241, 111725, 10.1016/j.rse.2020.111725, 2020.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C. A., III, Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C. J. L., and Forouzanfar, M. H.: Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015, The Lancet, 389, 1907-1918, 10.1016/S0140-6736(17)30505-6, 2017.

Colmer, J., Hardman, I., Shimshack, J., and Voorheis, J.: Disparities in PM2.5 air pollution in the United

States, Science, 369, 575, 10.1126/science.aaz9353, 2020.

Copat, C., Cristaldi, A., Fiore, M., Grasso, A., Zuccarello, P., Signorelli, S. S., Conti, G. O., and Ferrante, M.: The role of air pollution (PM and NO2) in COVID-19 spread and lethality: A systematic review, Environmental Research, 191, 110129, 10.1016/j.envres.2020.110129, 2020.

Dedoussi, I. C., Eastham, S. D., Monier, E., and Barrett, S. R. H.: Premature mortality related to United States cross-state air pollution, Nature, 578, 261-265, 10.1038/s41586-020-1983-8, 2020.

Gao, M., Gao, J., Zhu, B., Kumar, R., Lu, X., Song, S., Zhang, Y., Jia, B., Wang, P., Beig, G., Hu, J., Ying, Q., Zhang, H., Sherman, P., and McElroy, M. B.: Ozone pollution over China and India: seasonality and sources, Atmos. Chem. Phys., 20, 4399-4414, 10.5194/acp-20-4399-2020, 2020.

Grinsztajn, L., Oyallon, E., and Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data?, arXiv e-prints, arXiv:2207.08815, 2022.

Han, L., Zhao, J., Gao, Y., and Gu, Z.: Prediction and evaluation of spatial distributions of ozone and urban heat island using a machine learning modified land use regression method, Sustainable Cities and Society, 78, 103643, https://doi.org/10.1016/j.scs.2021.103643, 2022.

He, Q. and Huang, B.: Satellite-based mapping of daily high-resolution ground PM2.5 in China via space-time regression modeling, Remote Sensing of Environment, 206, 72-83, 10.1016/j.rse.2017.12.018, 2018.

Ialongo, I., Virta, H., Eskes, H., Hovila, J., and Douros, J.: Comparison of TROPOMI/Sentinel-5 Precursor NO2 observations with ground-based measurements in Helsinki, Atmos. Meas. Tech., 13, 205-218, 10.5194/amt-13-205-2020, 2020.

Ivey, C., Holmes, H., Shi, G., Balachandran, S., Hu, Y., and Russell, A. G.: Development of PM2.5 Source Profiles Using a Hybrid Chemical Transport-Receptor Modeling Approach, Environmental Science & Technology, 51, 13788-13796, 10.1021/acs.est.7b03781, 2017.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., and Kim, S.: Estimation of surface-level NO2 and O3 concentrations using TROPOMI data and machine learning over East Asia, Environmental Pollution, 288, 117711, https://doi.org/10.1016/j.envpol.2021.117711, 2021.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: a highly efficient gradient boosting decision tree, Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA2017.

Kim, M., Brunner, D., and Kuhlmann, G.: Importance of satellite observations for high-resolution mapping of near-surface NO2 by machine learning, Remote Sensing of Environment, 264, 10.1016/j.rse.2021.112573, 2021.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S.: Self-normalizing neural networks, Advances in neural information processing systems, 30, 2017.

Landrigan, P. J.: Air pollution and health, The Lancet Public Health, 2, e4-e5, 10.1016/S2468-2667(16)30023-8, 2017.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R.: Efficient backprop, in: Neural networks: Tricks of the trade, Springer, 9-48, 2012.

Li, R., Cui, L., Meng, Y., Zhao, Y., and Fu, H.: Satellite-based prediction of daily SO2 exposure across China using a high-quality random forest-spatiotemporal Kriging (RF-STK) model for health risk assessment, Atmospheric Environment, 208, 10-19, https://doi.org/10.1016/j.atmosenv.2019.03.029, 2019.

Li, T., Shen, H., Yuan, Q., Zhang, X., and Zhang, L.: Estimating Ground-Level PM2.5 by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach, Geophysical Research Letters, 44,

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

11,985-911,993, 10.1002/2017GL075710, 2017.

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., and Bi, J.: Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach, Environment International, 142, 105823, https://doi.org/10.1016/j.envint.2020.105823, 2020.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 56-67, 10.1038/s42256-019-0138-9, 2020.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., and Lee, S.-I.: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature Biomedical Engineering, 2, 749-760, 10.1038/s41551-018-0304-0, 2018.

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., and Reid, J. S.: Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm, J. Geophys. Res-Atmos., 116, 10.1029/2010JD014986, 2011.

MEEPRC: Bulletin on China's Ecological Environment(2019), 2020.

MEEPRC: Bulletin on China's Ecological Environment(2020) 2021.

Shen, G., Ru, M., Du, W., Zhu, X., Zhong, Q., Chen, Y., Shen, H., Yun, X., Meng, W., Liu, J., Cheng, H., Hu, J., Guan, D., and Tao, S.: Impacts of air pollutants from rural Chinese households under the rapid residential energy transition, Nature Communications, 10, 3405, 10.1038/s41467-019-11453-w, 2019.

Shwartz-Ziv, R. and Armon, A.: Tabular data: Deep learning is not all you need, Information Fusion, 81, 84-90, https://doi.org/10.1016/j.inffus.2021.11.011, 2022.

Silibello, C., Carlino, G., Stafoggia, M., Gariazzo, C., Finardi, S., Pepe, N., Radice, P., Forastiere, F., and Viegi, G.: Spatial-temporal prediction of ambient nitrogen dioxide and ozone levels over Italy using a Random Forest model for population exposure assessment, Air Quality, Atmosphere & Health, 14, 817-829, 10.1007/s11869-021-00981-4, 2021.

Wang, W., Liu, X., Bi, J., and Liu, Y.: A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology, Environment International, 158, 106917, https://doi.org/10.1016/j.envint.2021.106917, 2022.

Watson, G. L., Telesca, D., Reid, C. E., Pfister, G. G., and Jerrett, M.: Machine learning models accurately predict ozone exposure during wildfire events, Environmental Pollution, 254, 112792, https://doi.org/10.1016/j.envpol.2019.06.088, 2019.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., and Cribb, M.: Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach, Remote Sensing of Environment, 231, 10.1016/j.rse.2019.111221, 2019.

Wei, J., Li, Z., Li, K., Dickerson, R. R., Pinker, R. T., Wang, J., Liu, X., Sun, L., Xue, W., and Cribb, M.: Full-coverage mapping and spatiotemporal variations of ground-level ozone (O3) pollution from 2013 to 2020 across China, Remote Sensing of Environment, 270, 112775, https://doi.org/10.1016/j.rse.2021.112775, 2022.

Wen, X., Chen, W., Chen, B., Yang, C., Tu, G., and Cheng, T.: Does the prohibition on open burning of straw mitigate air pollution? An empirical study in Jilin Province of China in the post-harvest season, Journal of Environmental Management, 264, 110451, https://doi.org/10.1016/j.jenvman.2020.110451, 2020.

World Health, O.: WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization, Geneva2021.

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data, Environmental Science & Technology, 52, 13260-13269, 10.1021/acs.est.8b02917, 2018.

Xue, R., Wang, S., Li, D., Zou, Z., Chan, K. L., Valks, P., Saiz-Lopez, A., and Zhou, B.: Spatio-temporal variations in NO2 and SO2 over Shanghai and Chongming Eco-Island measured by Ozone Monitoring Instrument (OMI) during 2008–2017, Journal of Cleaner Production, 258, 120563, 10.1016/j.jclepro.2020.120563, 2020.

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., and Wang, W.: National-Scale Estimates of Ground-Level PM2.5 Concentration in China Using Geographically Weighted Regression Based on 3 km Resolution MODIS AOD, 10.3390/rs8030184, 2016.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., and Di, B.: Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment, Environmental Pollution, 233, 464-473, https://doi.org/10.1016/j.envpol.2017.10.029, 2018.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., and Zhang, M.: Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm, Atmospheric Environment, 155, 129-139, https://doi.org/10.1016/j.atmosenv.2017.02.023, 2017a.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., and Zhang, M.: Spatiotemporal prediction of continuous daily PM 2.5 concentrations across China using a spatially explicit machine learning algorithm, Atmospheric Environment, 155, 129-139, 10.1016/j.atmosenv.2017.02.023, 2017b.

Zhang, H., Di, B., Liu, D., Li, J., and Zhan, Y.: Spatiotemporal distributions of ambient SO2 across China based on satellite retrievals and ground observations: Substantial decrease in human exposure during 2013–2016, Environmental Research, 179, 108795, https://doi.org/10.1016/j.envres.2019.108795, 2019.

Zhang, T., Zhu, Z., Gong, W., Zhu, Z., Sun, K., Wang, L., Huang, Y., Mao, F., Shen, H., Li, Z., and Xu, K.: Estimation of ultrahigh resolution PM2.5 concentrations in urban areas using 160 m Gaofen-1 AOD retrievals, Remote Sensing of Environment, 216, 91-104, 10.1016/j.rse.2018.06.030, 2018a.

Zhang, X., Wang, Z., Cheng, M., Wu, X., Zhan, N., and Xu, J.: Long-term ambient SO2 concentration and its exposure risk across China inferred from OMI observations from 2005 to 2018, Atmospheric Research, 247, 105150, https://doi.org/10.1016/j.atmosres.2020.105150, 2021.

Zhang, Z., Wang, J., Hart, J. E., Laden, F., Zhao, C., Li, T., Zheng, P., Li, D., Ye, Z., and Chen, K.: National scale spatiotemporal land-use regression model for PM2.5, PM10 and NO2 concentration in China, Atmospheric Environment, 192, 48-54, https://doi.org/10.1016/j.atmosenv.2018.08.046, 2018b.

Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., and Zhang, W.: Robust prediction of hourly PM2.5 from meteorological data using LightGBM, National Science Review, 8, nwaa307, 10.1093/nsr/nwaa307, 2021.

Chi Y., Zhan Y., Wang Kai. & Ye H.. Spatial distribution of various air pollutants in China at 1 km(PM2.5 2015-01-01:2018-03-18). Zenodo. https://doi.org/10.5281/zenodo.7533813, 2023a.

Chi Y., Zhan Y., Wang Kai. & Ye H.. Spatial distribution of various air pollutants in China at 1 km(PM2.5 2018-03-19:2020-12-31). Zenodo. https://doi.org/10.5281/zenodo.7547774, 2023b.

Chi Y., Zhan Y., Wang Kai. & Ye H.. Spatial distribution of various air pollutants in China at 1 km (SO2 2015-01-01:2018-03-21). Zenodo. https://doi.org/10.5281/zenodo.7312179, 2023c.

Chi Y., Zhan Y., Wang Kai. & Ye H.. Spatial distribution of various air pollutants in China at 1 km

https://doi.org/10.5194/essd-2023-76
Preprint. Discussion started: 8 March 2023

(SO2 2018-03-21:2020-12-31). Zenodo. https://doi.org/10.5281/zenodo.7580714, 2023d.

Chi Y., Zhan Y., Wang Kai. & Ye H.. Spatial distribution of various air pollutants in China at 1 km(O3 2015-01-01:2018-02-28). Zenodo. https://doi.org/10.5281/zenodo.7580720, 2023e.

Chi Y., Zhan Y., Wang Kai. & Ye H.. Spatial distribution of various air pollutants in China at 1 km(O3 2018-03-01:2020-12-31). Zenodo. https://doi.org/10.5281/zenodo.7580726, 2023f.