Environmental Pollution 276 (2021) 116635

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Random forest model based fine scale spatiotemporal O_3 trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017☆

Runmei Ma [a, 1], Jie Ban [a, 1], Qing Wang [a], Yayi Zhang [a, b], Yang Yang [c], Mike Z. He [d], Shenshen Li [e], Wenjiao Shi [f], Tiantian Li [a, *]

[a] China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing, 100021, China
[b] Jiangsu Ocean University, Jiangsu, 222000, China
[c] Institute of Urban Meteorology, China Meteorological Administration, Beijing, 100089, China
[d] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York NY, 10029, USA
[e] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute (AIR), Chinese Academy of Sciences, Beijing, 100101, China
[f] Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China

ARTICLE INFO

Article history:
Received 19 October 2020
Received in revised form
12 January 2021
Accepted 29 January 2021
Available online 2 February 2021

Keywords:
Ambient ozone
Random forest model
Simulation

ABSTRACT

Ambient ozone (O_3) concentrations have shown an upward trend in China and its health hazards have also been recognized in recent years. High-resolution exposure data based on statistical models are needed. Our study aimed to build high-performance random forest (RF) models based on training data from 2013 to 2017 in the Beijing-Tianjin-Hebei (BTH) region in China at a 0.01 ° × 0.01 ° resolution, and estimated daily maximum 8h average O_3 (O_3-8hmax) concentration, daily average O_3 (O_3-mean) concentration, and daily maximum 1h O_3 (O_3-1hmax) concentration from 2010 to 2017. Model features included meteorological variables, chemical transport model output variables, geographic variables, and population data. The test-R^2 of sample-based O_3-8hmax, O_3-mean and O_3-1hmax models were all greater than 0.80, while the R^2 of site-based and date-based model were 0.68–0.87. From 2010 to 2017, O_3-8hmax, O_3-mean, and O_3-1hmax concentrations in the BTH region increased by 4.18 μg/m^3, 0.11 μg/m^3, and 4.71 μg/m^3, especially in more developed regions. Due to the influence of weather conditions, which showed high contribution to the model, the long-term spatial distribution of O_3 concentrations indicated a similar pattern as altitude, where high concentration levels were distributed in regions with higher altitude.

1. Introduction

As a major secondary air pollutant, ambient ozone (O_3) has been a growing concern in public health. According to the 2017 Global Burden of Disease (GBD) study, 233,638 (95% uncertainty interval: 90,109–385,303) premature deaths were attributable to ambient O_3 exposure globally in 2016 (Dicker et al., 2018). Both short- and long-term exposure to ambient O_3 have been linked to total mortality, cardiovascular diseases, and respiratory diseases (Yin et al., 2017; Bell et al., 2004; Wong et al., 2008; Peng et al., 2013; Di et al., 2017a; Yang et al., 2017), as well as biomarker levels such as blood glucose (Yang et al., 2018), blood pressure (Cole-Hunter et al., 2018), and inflammation factors (Lee et al., 2018). Although a strict pollution control measure, the *Air Pollution Prevention and Control Action Plan*, was enacted in China in 2013, ambient O_3 was the only pollutant with an increasing trend in both concentration levels and number of days exceeding the air quality standard: the O_3 maximum 8-h average concentration in China in 2018 was 151 μg/m^3, the proportion of days exceeding the standard is 8.4%, which is 0.8% higher than 2017 (Ministry of Ecology and Environment of the People's Republic of China, 2019). From a public health perspective, accurate ambient O_3 exposure assessment is critically needed to address the health effects of short- and long-term exposure to O_3 in environmental epidemiologic research and policy recommendations.

☆ This paper has been recommended for acceptance by Pavlos Kassomenos
* Corresponding author.
E-mail address: litiantian@nieh.chinacdc.cn (T. Li).
[1] Runmei Ma and Jie Ban are the co-first authors.

https://doi.org/10.1016/j.envpol.2021.116635
0269-7491/

Nomenclature

O_3	Ambient ozone
GBD	Global Burden of Disease
BTH	Beijing-Tianjin-Hebei
LUR	Land use regression
RF	Random forest
O_3-8hmax	Daily maximum 8h average O_3
O_3-mean	Daily average O_3
O_3-1hmax	Daily maximum 1h O_3
CV	Cross-validation
MEIC	Multi-resolution Emission Inventory for China
OMI	Ozone monitoring instrument
RESDC	Resource and Environmental Science Data Center
NDVI	Normalized Difference Vegetation Index
RMSE	Root mean square error
MAE	Mean absolute error
XGBOOST	eXtreme Gradient Boosting

However, a lack of fine-scale spatial-temporal measurements of ambient O_3 concentrations during the past decades has posed a gap in linking O_3 exposure and health outcomes, especially in heavily polluted regions in China. On one hand, it was not until 2013 that 74 Chinese cities officially launched routine air quality assessment on ambient O_3, with 1436 national monitoring sites (China National Environmen, 2017) across China established by the end of 2014. While more monitors are being added, existing monitors are sparse and introduces potential uncertainties in exposure assessment (Zeger et al., 2000). On the other hand, although there are a few recently published works provided O_3 modeling data in China at the national scale (Liu et al., 2020; Zhan et al., 2018), the following issues still remain to be addressed: first, only single metrics of O_3 concentration was discussed, while there exists different metrics that are potentially meaningful for the purposes of epidemiologic studies; second, generally, the resolution of existing modeled data is at the 0.1° grid level, and finer resolution data may be better for eliminating exposure measurement errors in pollution hotspot areas, such as the Beijing-Tianjin-Hebei (BTH) region.

Previously, we conducted a review (Ma et al., 2020) on ambient O_3 simulations in statistical models to confirm the feasibility of simulation studies. In general, land use regression (LUR) models, random forest (RF) models, and artificial neural network models have been used, while machine learning showed higher model performance in large-scale, long duration, and high-spatial-temporal resolution research. Furthermore, in comparisons with other modeling method, RF models demonstrated superiorities in minimizing error (Ren et al., 2020).

Therefore, our study aims to simulate three O_3 metrics, including daily maximum 8h average O_3 (O_3-8hmax) concentration, daily average O_3 (O_3-mean) concentration, and daily maximum 1h O_3 (O_3-1hmax) concentration in the BTH region from 2010 to 2017 using RF models at the 0.01 ° × 0.01 ° resolution. First, we introduce the entire process of establishing a multi-parameter RF model based on sample-based data division and 10-fold cross-validation (CV). Second, we test model robustness by building model through site-based and date-based data division and 10-fold CV. Finally, we analyze the spatial and temporal trends of ambient O_3 concentrations in the BTH region from 2010 to 2017.

2. Methods

2.1. Study area and study period

The study area is located in the BTH region, which is one of the economic centers of China with more than 100 million people (National Bureau of Statistics, 2020). It is also one of the key areas for air pollution control on the national scale, and one of the areas of greatest concern for research on air pollution and health in China (Zhao et al., 2020). On the basis of the province borders, an extra 0.5° spatial range was extended to adequately capture the O_3 exposure in a particular province (113.45°E–119.85°E and 36.03°N-42.62°N) (Fig. 1). The study period was from January 1st, 2010 to December 31st, 2017, including 2992 days. Because ambient O_3 monitoring officially started in 2013, the historical period in 2010–2012 was simulated based on models using O_3 measurements from 2013 to 2017.

2.2. Model features

The variables included in our model are determined based on previous simulation studies as well as data availability (Ma et al., 2020). Multi-resolution Emission Inventory for China (MEIC) and ozone column concentration from the ozone monitoring instrument (OMI) in Aura were excluded, because of the low spatial-temporal resolution and high missing value (Figure S1). In previous modeling tests, these two variables also showed low contribution. A full list of model variables is shown in Table S1.

2.2.1. Ambient O_3 measurement data

Hourly ambient O_3 concentrations from 2013 to 2017 (January 1st, 2013 to December 31st, 2017) were obtained from the China National Environmental Monitoring Centre (http://www.cnemc.cn/). The qualified stations were defined as sites with a missing rate of less than 25% throughout the study period. The exposure indicators used in our study included three metrics: O_3-8hmax, O_3-mean, and O_3-1hmax concentration. On this basis, the daily level indicator can be calculated only if the original hour level data of each site is $\geq$ 6 h. A total of 95 sites were included in our study (Fig. 1).

2.2.2. Meteorological variables

Meteorological data from 2010 to 2017 were taken from the ERA-Interim reanalysis data of the European Center for Medium-Range Weather Forecasts (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/). The spatial and temporal resolution of these variables were 0.125 ° × 0.125 ° and 6/12 h, respectively. 45 types of meteorological variables were included in our study after considering the formation mechanism and previous simulations studies of ambient O_3. Lag effects, such as the ones we evaluated in our previous $PM_{2.5}$ modeling work (Zhao et al., 2020), were also considered by including the previous two days' values (lag1 and lag2) of these variables.

2.2.3. The chemical transport model output

To simulate the formation and dispersion of ambient O_3, ambient O_3 concentrations in 2010–2016 came from outputs from the GEOS-Chem chemical transport model (Weagle et al., 2018), which incorporates meteorological conditions, emissions outputs, and chemical reactions. The spatial and temporal resolution of the model was 2 ° × 2.5 ° and 2 h, respectively. The original data are divided into 37 layers based on altitude. This study extracts the near-surface layer raw data and calculates its daily average.

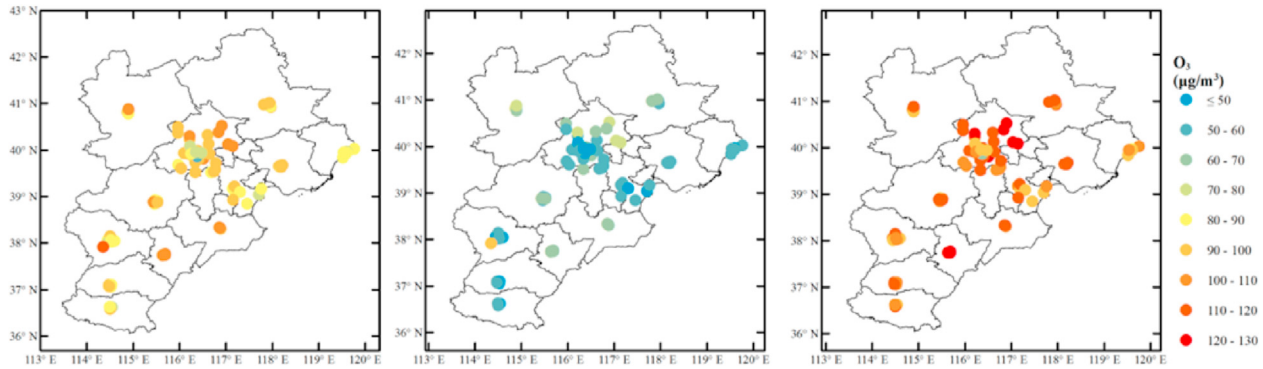

Fig. 1. The monitoring sites and ambient O_3 level from 2013 to 2017. (The points in figures represent the monitoring sites; From left to right are O_3-8hmax, O_3-mean and O_3-1hmax level, respectively).

2.2.4. Geographic variables

Annual land use data from 2010 to 2017 were downloaded from the European Space Agency Climate Change Initiative (http://maps.elie.ucl.ac.be/CCI/viewer/) website. Three types of land use data, including natural vegetation coverage, urban coverage, and farmland coverage, were extracted.

The road network data of 2016 were published by the Resource and Environmental Science Data Center (RESDC) of the Chinese Academy of Sciences (http://www.resdc.cn). In our study, national highway length, highway length, provincial highway length, railway length, county road length, and country road length were included.

Normalized Difference Vegetation Index (NDVI) data from 2010 to 2016 were taken from the International Scientific and Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences (http://www.gscloud.cn). The spatial and temporal resolution of the MODND1D product was monthly at 500 m × 500 m.

Elevation data with a spatial resolution of 1 km × 1 km for 2010 came from the RESDC of the Chinese Academy of Sciences (http://www.resdc.cn).

2.2.5. Population data

National population data for 2010 were collected from the RESDC of Chinese Academy of Sciences (http://www.resdc.cn), and were distributed on a 1 km × 1 km grid.

2.2.6. Dummy parameters

Considering the varied time pattern of ambient O_3 concentrations, we set time dummy variables including season (spring, summer, fall and winter) and month (January to December).

2.3. Data process

We built a 0.01 ° × 0.01 ° standard grid of the BTH region for data integration. O_3 measurements were converted into gridded data and assigned to grid cell based on site coordinates using tools from the *sp* and *raster* packages in R. GEOS-Chem model outputs were processed to 0.01 ° × 0.01 ° by inverse distance weighted interpolation using *idw* package in Python. Meteorology data, land use data, population data, elevation data and NDVI are evenly distributed, and downscaled using *Extract Values to Points* with bilinear interpolation in ArcGIS 10.2. Road network data were processed through the *intersect* and *dissolve* function into grids in ArcGIS 10.2. For long term parameters, each day within a month or year was assigned the corresponding month or year level value. A dataset including all parameters, dummy variables, and ambient O_3 measurements was prepared for model development in the next step.

2.4. Model development and validation

The RF model was developed by including multiple decision trees that were generated by the bagging ensemble method (Breiman, 2001). The sample-based division method randomly divides data into a training set and test set, where the training set included 90% of the data and the test set included 10% of the data. The training set was used to build the model based on a 10-fold cross-validation. The process was performed by randomly dividing the training set into 10 subsets, where nine subsets were used to build the model and 1 subset was used to validate the model performance; The analysis was performed ten times, and the average of the 10 runs was used as the final result. The model was built using following formula:

$$Y_{i,j} = f\left(METE_{i,j},\ lag1METE_{ij}, lag2METE_{i,j}, GEOS_{i,j}, LD_j, ROAD_j, NDVI_j, ELE_j, POP_j, SEASON_{i,j}, MON_{i,j}\right) \quad (1)$$

Where $Y_{i,j}$ was the O_3 concentration on day i in grid cell j; METE, lag 1 $METE_{i,j}$ and lag2 $METE_{i,j}$ were meteorological variables and its one-day lag and two-day lag values on day i in grid cell j; $GEOS_{i,j}$ was the GEOS-Chem model output on day i in grid cell j; LD_j, $ROAD_j$, $NDVI_j$, ELE_j and POP_j were land use coverage, length of the road, NDVI product, elevation, and population in grid cell j, respectively; and $SEASON_{i,j}$ and $MON_{i,j}$ were the season and month of year, respectively.

To avoid over-fitting, as well as potential poor generalizability of model, the model parameter, maximum depth (the maximum depth of the tree), needed to be adjusted. The *Grid Search* method in Python was used. The error rate was calculated by employing predictions from out-of-bag samples. Based on the results of *Grid*

Search, we set the maximum depth as 47, 35 and 47 for O_3-8hmax model, O_3-mean model and O_3-1hmax model, respectively, to achieve high prediction accuracy in the experiments. We set n_estimators (the number of trees) as 200 for all three models. The R^2, root mean square error (RMSE) and mean absolute error (MAE) for different years, months, provinces, municipalities and counties were calculated to test model performance.

As a supplement to the above sample-based division method, we also implemented site-based and date-based division methods to verify the stability of the model on spatial and temporal scales. Site-based division method means that monitoring data from 90% of the stations that were randomly selected were used as the training set, while monitoring data from the remaining stations were used as the testing set. Date-based division method means that monitoring data for ten days in January, April, July and October of each year were randomly selected as the testing set, for which 200 days in total were selected; monitoring data for the remaining days were used as the training set. The entire modeling process was the same as that of the sample-based model.

2.5. Spatiotemporal simulation of ambient O_3 concentration

Using the final sample-based model, the daily O_3-8hmax, O_3-mean, and O_3-1hmax concentration in the BTH region from 2010 to 2017 were simulated. The yearly and seasonal average concentrations were calculated. The spatial and temporal trends of ambient O_3 were analyzed combined with China's environmental protection policies and regulations.

All modeling and simulation work were performed in Python 2.7 based on *scikit-learn* package. The workflow of estimating the spatiotemporal ambient O_3 concentrations in this study is shown in Fig. 2.

3. Results

3.1. Ambient summary of O_3 concentration and parameters

The mean ± standard deviation of O_3-8hmax, O_3-mean, and O_3-1hmax concentrations from 2013 to 2017 were 92.58 ± 60.04 μg/m^3, 58.39 ± 38.11 μg/m^3 and 110.75 ± 68.95 μg/m^3, respectively (Fig. 1). The descriptive statistics of the variables are shown in Table S2.

3.2. Feature importance

The modeling importance of the top ten variables are shown in Table 1, and the full results are shown in Table S3. During the O_3-8hmax and O_3-1hmax modeling, 2-m temperature showed the highest importance, which accounted for 47% and 49% of all relative importance; while GEOS-Chem outputs accounted for 42% during the O_3-mean modeling. The meteorological variables showed high influence to daily ambient O_3 concentrations: downward surface solar radiation, V wind component, and low cloud cover were all among the most important variables in all three models. Altitude, which is a long-term parameter, also showed high importance to ambient O_3 simulations in our models.

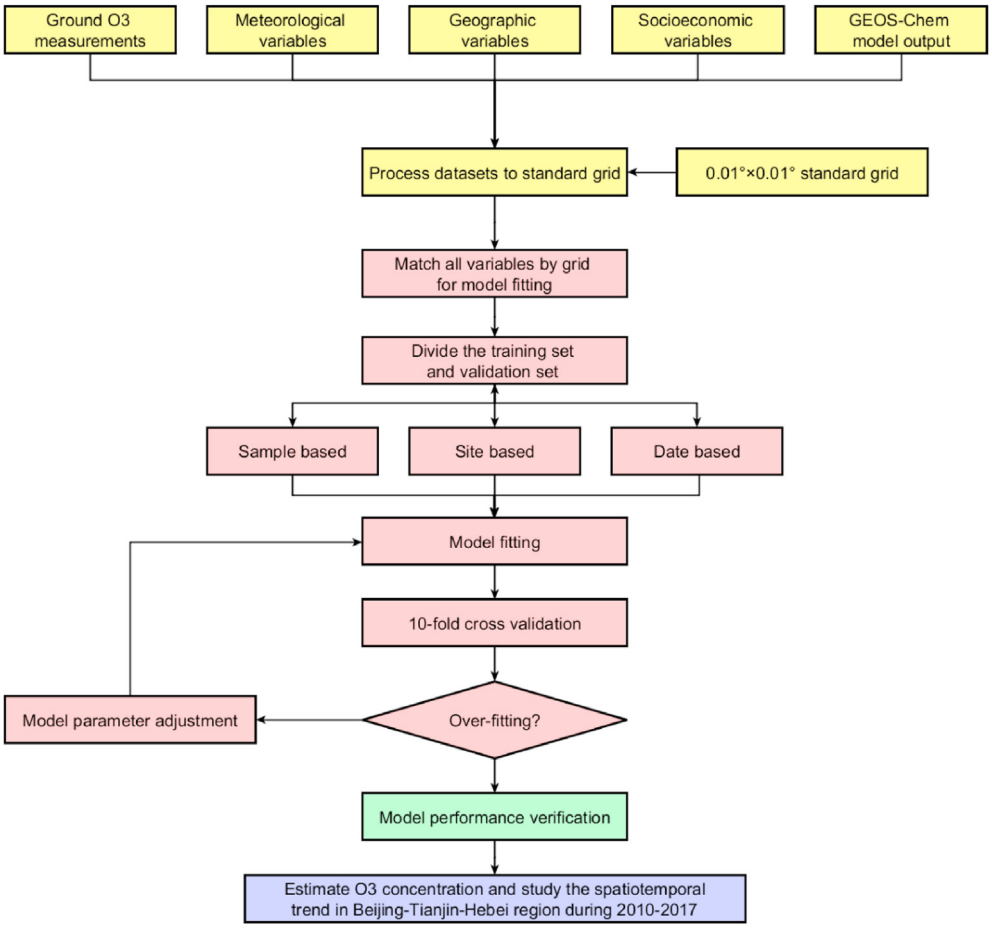

Fig. 2. The workflow of ambient O_3 concentrations simulation in this study.

Table 1
The importance value of top ten variables in O_3-8hmax, O_3-mean and O_3-1hmax model.

Sort	O_3-8hmax		O_3-mean		O_3-1hmax	
	Variables	Value	Variables	Value	Variables	Value
1	2-m temperature	47%	GEOS-Chem	42%	2-m temperature	49%
2	Downward surface solar radiation	10%	Downward surface solar radiation	8%	Downward surface solar radiation	7%
3	10-m V wind component	5%	10-m V wind component	5%	10-m V wind component	5%
4	GEOS-Chem	4%	2-m temperature	5%	GEOS-Chem	2%
5	Low cloud cover	2%	10-m V wind component lag1	2%	Low cloud cover	2%
6	10-m V wind component lag1	2%	Dem	2%	10-m V wind component lag1	2%
7	Dem	1%	Low cloud cover lag1	2%	Dem	2%
8	Medium cloud cover	1%	month	2%	Boundary layer height lag1	1%
9	2-m dewpoint temperature	1%	Downward surface solar radiation lag1	1%	Medium cloud cover	1%
10	Boundary layer height lag1	1%	Boundary layer height lag1	1%	Downward surface solar radiation lag1	1%

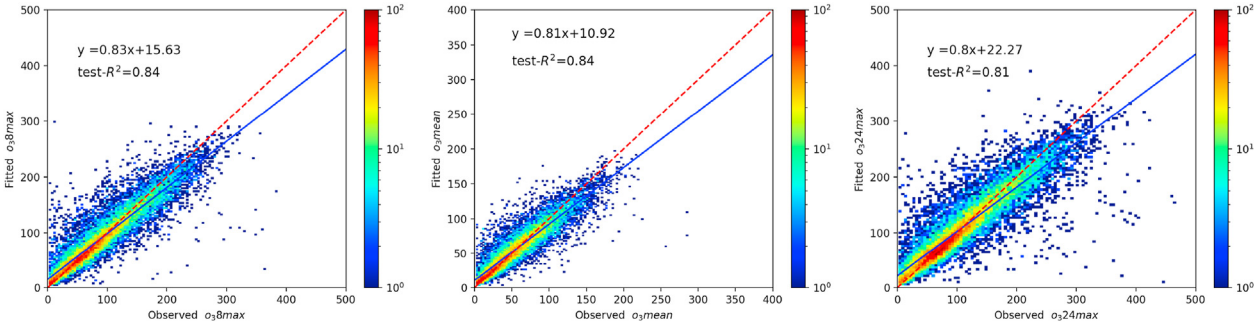

Fig. 3. Density scatter plots of the ambient O_3 model test validation results. (From left to right are O_3-8hmax, O_3-mean, and O_3-1hmax level, respectively).

3.3. Model performance

The three models of daily ambient O_3 concentration showed high performance, with R^2 values of 0.84 for O_3-8hmax, 0.84 for O_3-mean, and 0.81 for O_3-1hmax, respectively (Fig. 3). We evaluated model performance at the provincial and municipal levels (Table S4). At the provincial level, the model performance of Beijing was higher than those in Tianjin and Hebei in all three models. At the municipal level, R^2 values ranged from 0.63 to 0.92, and Hengshui, Langfang, Tangshan had the best predictions out of 11 cities.

The yearly and seasonal R^2s were also calculated (Figure S2 and Figure S3): as time progressed, the R^2 of each year from 2013 to 2017 also continued to increase, which may be related to the growth of China's monitoring network construction. All three models achieved the best model performance in autumn; O_3-8hmax and O_3-mean models had relatively poorer performance in the summer, while O_3-1hmax model performed relatively poorly in the winter.

The stabilities of models were tested both spatially and temporally through site-based and date-based division models (Figure S4 and Figure S5). From the spatial verification results, the R^2 of the O_3-8hmax model is 0.87, the R^2 of the O_3-mean model is 0.79, and the R^2 of the O_3-1hmax model is 0.84; From the temporal verification results, the R^2 of the O_3-8hmax model is 0.71, the R^2 of the O_3-mean model is 0.70, and the R^2 of the O_3-1hmax model is 0.68.

3.4. Spatial and temporal trends of ambient O_3 concentration in BTH region from 2010 to 2017

The modeling results of daily O_3-8hmax, O_3-mean, and O_3-1hmax with high spatial-temporal resolution in the BTH region from 2010 to 2017 were shown in Figure S6. The annual concentrations were shown in Figs. 4-6. Ambient O_3 concentrations were higher in the northern and western regions of the study area, while the eastern and southern regions had lower concentrations. This is consistent with the distribution of altitude in the study area: the terrain in the BTH region is high in the northwest and low in the southeast.

Ambient O_3 concentrations from 2010 to 2017 showed an overall upward trend, especially since 2016 (Table S5). 4.18 μg/m^3, 0.11 μg/m^3 and 4.71 μg/m^3 increase were found comparing 2017 to 2010 in O_3-8hmax, O_3-mean, and O_3-1hmax, separately. The areas with increasing concentration levels are mainly concentrated in the economically developed areas in the central and southern parts of the BTH region (Fig. 7). The modeling ambient O_3 concentration values show a clear seasonal trend: the O_3 levels are highest in the summer, followed by spring and autumn, and lowest in the winter, which is consistent with the formation mechanism of ambient O_3 (Figure S7).

4. Discussion

Based on high spatial resolution (0.01 ° × 0.01 °) models, our study has a clearer understanding of the temporal and spatial distribution of O_3 concentration in the BTH region from 2010 to 2017. Using a list of suitable variables, including meteorological variables, chemical transport model output, geographic variables, and population variables, our model achieved high performance, with R^2s for three indicators including O_3-8hmax, O_3-mean, and O_3-1hmax model all higher than 0.80. The 2010–2017 O_3 concentrations in the BTH area showed an overall increasing trend, especially since 2016.

Our model has achieved higher model performance at a high spatiotemporal resolution levels than those of other comparable studies. With a 0.1 ° × 0.1 ° resolution, Zhan et al. (2018) and Liu et al. (2020) assessed daily ambient O_3 concentrations in China

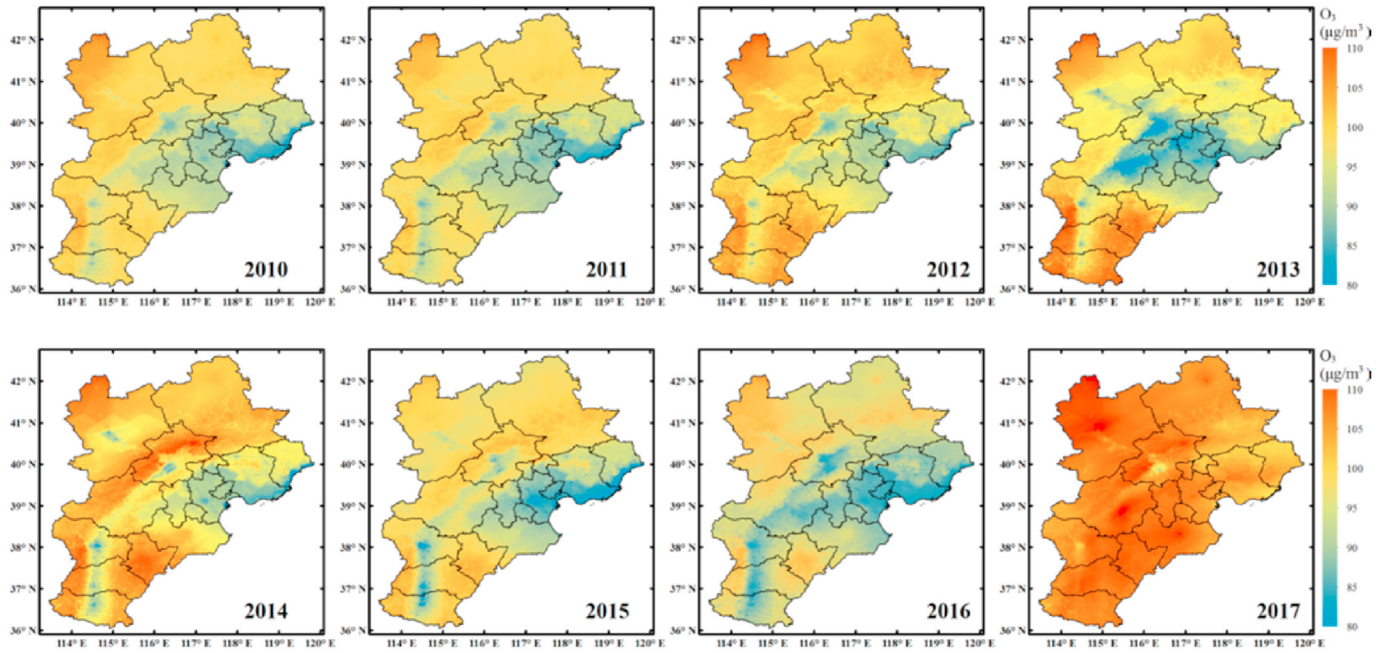

Fig. 4. Annual average simulation results of O_3-8hmax from 2010 to 2017 in the study area.

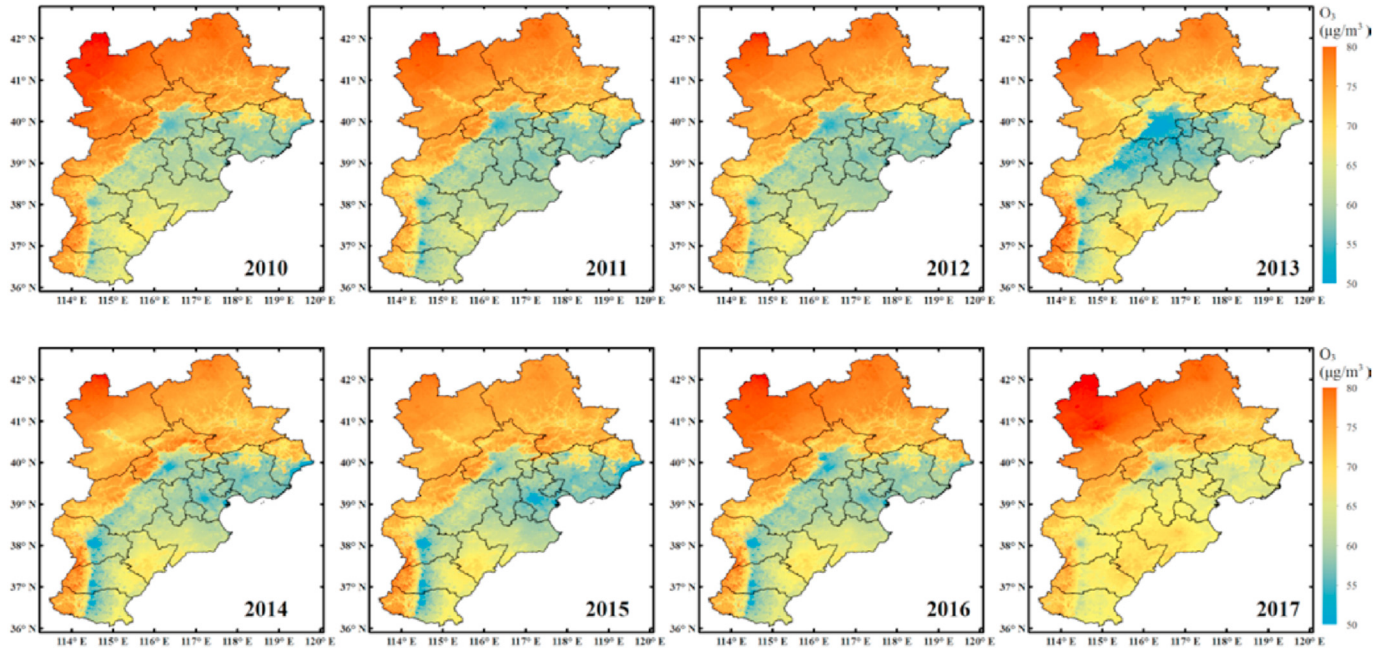

Fig. 5. Annual average simulation results of O_3-mean from 2010 to 2017 in the study area.

using a RF model and an eXtreme Gradient Boosting (XGBOOST) algorithm, separately, and the R^2 value were 0.61–0.78. Huang et al. (2017) modeled annual average O_3 concentrations in Nanjing, China in 2013 at a 100 m × 100 m resolution using LUR models, and the R^2 value was 0.65. Compared with studies in other countries, our model still showed comparable or even better performance: Ren et al. (2020) conducted 13 linear and non-linear models in the United States in 2011, and found better performance in RF and XGBOOST, with R^2 values of 0.84 and 0.81. The R^2 of a neural network model of daily concentrations with a 1 km × 1 km resolution in the United States (Di et al., 2017b) was 0.74–0.80; the R^2 of a LUR model with higher spatial resolution in six metropolitan area, U.S. (Wang et al., 2015), Los Angeles Basin, U.S. (Wang et al., 2016), and Augsburg, Germany (Wolf et al., 2017) were slightly higher than that of our study, which may due to the coarser time resolution in these studies (two-week and annual averages). On one hand, various model features can capture the trend of ambient O_3 concentration more comprehensively. Daily meteorological variables were included in the form of the day, lag1, and lag2, which fully considers the lagging effect of meteorological variables on O_3. Furthermore, GEOS-Chem model outputs were introduced in this study to make up for the shortcomings of pure statistical models

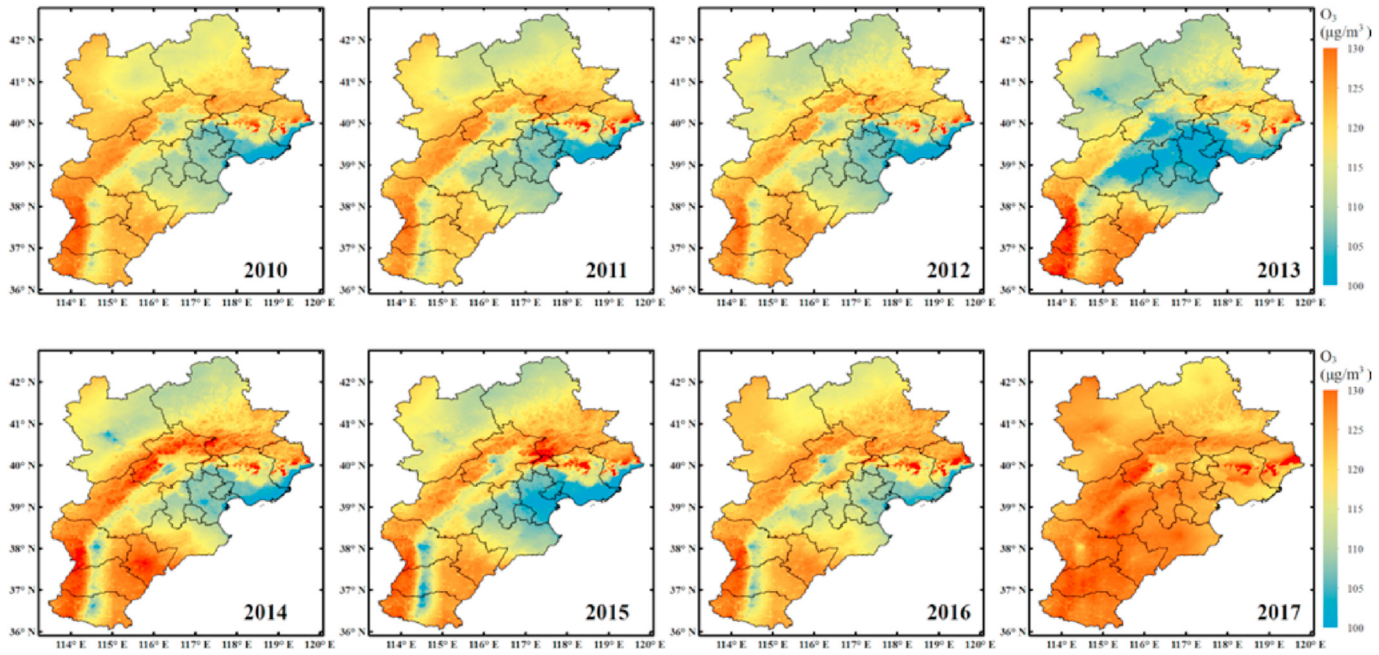

Fig. 6. Annual average simulation results of O_3-1hmax from 2010 to 2017 in the study area.

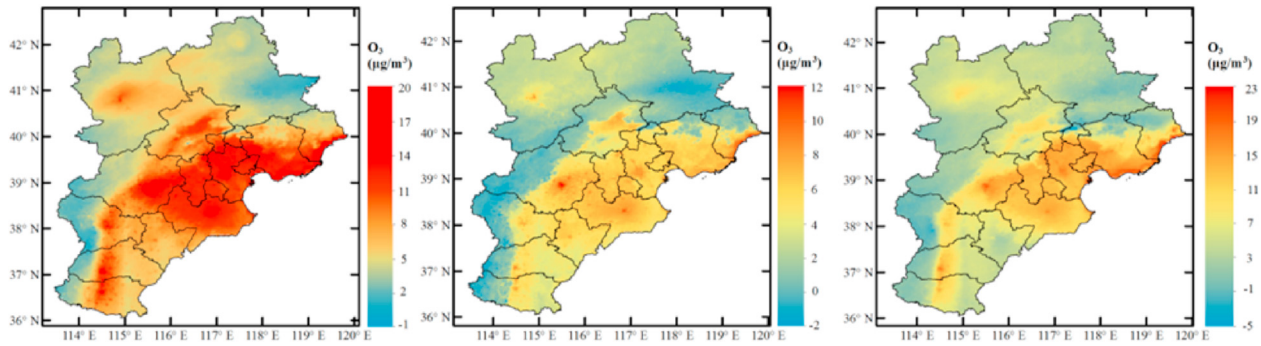

Fig. 7. The difference between 2010 and 2017 annual average O_3 concentrations in the study area (From left to right are O_3-8hmax, O_3-mean, and O_3-1hmax).

that do not consider chemical processes. On the other hand, the random forest model built nonlinear relationship between multiple features and O_3 concentration. Compared with linear model like LUR models, the RF model showed higher accuracy (Ren et al., 2020); we can also see the superiority of machine learning algorithms represented by RF in large-scale daily simulation studies (Ma et al., 2020; Hoek et al., 2008).

The inevitable uncertainty of historical simulations has been pointed out in previous studies (Zhan et al., 2018; Ma et al., 2016). The increased yearly R^2 indicates that the model performance is relatively poor when there is less monitoring data in the early years; the lower R^2 in the summer and winter showed our modeling work has limitations in capturing extreme values of ambient O_3. However, our model was established based on years of monitoring data and variables (2013–2017), which reduces the uncertainty. Through selecting specific dates that can represent typical climatic conditions as the test set for temporal verification, the stable results also confirmed the credibility of historical simulated datasets in some content. The limitations of forming grid exposure were mainly focused on the density of monitoring stations (Wang et al., 2015; Adam-Poupart et al., 2014). The stable spatial verification results showed that under the existing site distribution and spatial resolution settings, the grid exposures obtained were relatively credible.

In our study, we found meteorological variables, especially temperature and solar radiation, showing high sensitivity to ambient O_3. This is consistent with the formation mechanism of ambient O_3 and previous studies (Zhan et al., 2018; Son et al., 2018). Our study also found lagging effects of the V wind component, boundary layer height, and downward surface solar radiation. The weather conditions of the previous day have an impact on the ozone concentration of the day by affecting the diffusion and elimination of pollutants. Combined with the mechanism of ambient ozone formation, at higher emission levels, it is necessary to further prevent the health hazards caused by possible exposure to high levels of ambient O_3 in high temperature, no wind, and high solar radiation.

In addition, the GEOS-Chem output considers atmospheric chemical processes, which have been proven to improve model performance (Wang et al., 2016; De Hoogh et al., 2018). Elevation is the only long-term variable with high variable importance ranking. It is closely related to meteorological changes such as solar

radiation, wind direction, wind pressure, and boundary layer height; topographic differences also lead to changes in thermal and dynamic effects. Its importance has also been proved in previous studies (Wang et al., 2016; Beelen et al., 2009). Unlike some previous studies (Wang et al., 2016; Beelen et al., 2009), land use variables did not appear in the forefront of the importance ranking in our study. This may be due to its difficulty in providing information about temporal variation (Di et al., 2017b). Compared with varied meteorological factors and emissions of pollutant, the influence of land use variables is relatively small (Zhan et al., 2018) and their inclusion leads to a decrease of temporal R^2 (Di et al., 2017b). It's difficult to explain the wider dynamic range of ambient O_3 in China with land use variables. Therefore, meteorological factors and emissions information should be further considered.

Our model found difference in variables importance between three exposure indicators. Both the O_3-8hmax and O_3-1hmax models emphasize the contribution of temperature and solar radiation, while the O_3-mean model emphasizes the importance of the output of the GEOS-Chem model. There are two possible reasons. First, as mentioned previously, the GEOS-Chem output could fully describe the formation and dissipation process of ambient O_3; Second, GEOS-Chem output indicated higher importance in O_3-mean model may be due to the fact that they have the same timescale. Because the O_3-mean value is affected by the elimination effect during night times; compared with O_3-8hmax and O_3-1hmax, which reflect the peak level of O_3, the daily average GEOS-Chem model output is more suitable for O_3-mean. As an air pollutant with significant variations during the course of a day, modeling work of O_3 with higher temporal resolution is needed.

We found consistency between ambient O_3 levels and altitude, and O_3 level is higher in high altitude areas, which is consistent with a previous study (Di et al., 2017b). This may be due to the invasion of natural source O_3 from the stratosphere, which can produce some transient peak O_3 concentrations at the ground level (Davies and Schuepbach, 1994). For areas with relatively flat terrain on the southeastern side of the BTH region, the differences in spatial distribution are mainly due to the human-sourced O_3, which is consistent with the distribution characteristics of local transportation and petrochemical and coal-fired power plants.

The simulated concentrations based on our model supports us to initially analyze the long-term trend of ozone concentration in 2010–2017. We found that before the implementation of the *Air Pollution Prevention and Control Action Plan* in China in 2013, O_3 was already at a high level in China compared to other countries (Bell et al., 2004; Di et al., 2017a). After the implementation of strict control measures, a decline was found in PM_{10}, $PM_{2.5}$, and SO_2 concentration levels, but not O_3 concentrations (Ministry of Ecology and Environment of the People's Republic of China, 2019), especially in the more developed regions. VOC emissions that have not been effectively controlled yet (Chen, 2017) may be one of the main reasons: due to the complex relationship between NOx and VOCs for ozone generation, the initial results of NOx treatment (Chen, 2017) have made it difficult for free radicals produced by VOCs to be consumed. Under the action of radiation, it has promoted the production of secondary pollutants. It has been confirmed during the COVID-19 lockdown in China (Huang et al., 2020), where large decreases in NOx emissions were seen from decreased transportation, ozone concentrations actually increased. Therefore, effective control strategies about collaborative management of VOCs and NOx in key industries and areas are needed when facing the strengthened coordinated development (Chen, 2017) and the low atmospheric dispersion conditions (Feng et al., 2017) in the BTH region. In addition, due to the consistent trend of the highest concentration of the three indicators in summer, future early warning work on severely polluted weather events in summer, especially with adverse meteorological conditions (high temperature, no wind, and high solar radiation), is also needed.

Our study has great potential relevance for future research. The O_3-8hamx, O_3-mean, and O_3-1hmax concentration in the historical period of the BTH region with high spatial resolution can be further traced, which could fill the existing gaps in exposure before 2013 and reduce exposure measurement error. The modeling dataset can be further applied in epidemiologic studies in the future.

Our study has some uncertainties. The major limitation is the uncertainty of monitoring data in the spatial and temporal scale. For the locations where monitoring data are not available, for example, northwest of the BTH region where monitoring sites are sparsely distributed, there may not be enough samples to capture the accurate association between O_3 concentration and model variables. Although we used the site-based division method to test the model performance, uncertainty remains. For the temporal perspective, although we have implemented a date-based division in our model, the validation of the historical period before 2013 could not be conducted. Second, the spatial resolution of each variable in the model is different, and the interpolation process inevitably introduces errors. Third, O_3 precursor concentrations including NOx and VOCs, and emission inventory with high spatial and temporal resolution were lacking due to data availability; traffic flow is also an important feature reflecting the variation of precursors, and its temporal and spatial variation trend plays an important role in capturing the characteristics of ambient O_3 concentration. Previous studies have suggested the importance of these parameters, so they should be considered in future studies when possible.

5. Conclusion

We built high-performance random forest models using meteorological features, geographical features, socioeconomic features for daily O_3-8hmax, O_3-mean, and O_3-1hmax concentration in the BTH region. The high-resolution exposure data is already available (https://cepht.niehs.cn:8282/developSDS.html) for future environmental epidemiologic studies. From 2010 to 2017, O_3-8hmax, O_3-mean, and O_3-1hmax concentrations in the BTH area increased, especially in more developed regions. Strengthening control of key industries and regions, and synergistic control of NOx and VOCs are the key points of ambient O_3 control in the BTH region. The public should protect themselves especially on days when ambient ozone pollution is likely to occur, such as high temperatures, high solar radiation, and heavy pollution.

Author contributions

Runmei Ma and **Jie Ban**: Software, Investigation, Validation, Formal analysis, Data curation, Writing - original draft. **Qing Wang and Yayi Zhang**: Visualization, Writing - Original Draft. **Yang Yang, Mike Z He, Shenshen Li and Wenjiao Shi**: Methodology, Writing - Review & Editing. **Tiantian Li**: Conceptualization, Methodology, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by grants from National Key Research and Development Program of China (Grant: 2017YFC0211706) and National Natural Science Foundation of China (Grant No. 92043301 and No. 41701234).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2021.116635.

References

Adam-Poupart, A., Brand, A., Fournier, M., et al., 2014. Spatiotemporal modeling of ozone levels in quebec (Canada): a comparison of kriging, land-use regression (LUR), and combined bayesian maximum entropy–LUR approaches. Environ. Health Perspect. 122 (9), 970.

Beelen, R., Hoek, G., Pebesma, E., et al., 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Sci. Total Environ. 407 (6), 1852–1867.

Bell, M.L., McDermott, A., Zeger, S.L., et al., 2004. Ozone and short-term mortality in 95 US urban communities, 1987-2000. J. Am. Med. Assoc. 292 (19), 2372–2378.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chen, Jining, 2017. Taking improving the ambient air quality as the core, we will resolutely win the three-year plan on defending the blue sky. Current Affairs Report 45–59, 003.

China National Environmental Monitoring Centre, 2017. National Environmental Air Quality Monitoring Network. http://www.cnemc.cn/zzjj/jcwl/dqjcwl/201711/t20171108_645109.shtml.

Cole-Hunter, T., de Nazelle, A., Donaire-Gonzalez, D., et al., 2018. Estimated effects of air pollution and space-time-activity on cardiopulmonary outcomes in healthy adults: a repeated measures study. Environ. Int. 111, 247–259.

Davies, T.D., Schuepbach, E., 1994. Episodes of high ozone concentrations at the earth's surface resulting from transport down from the upper troposphere/lower stratosphere: a review and case studies. Atmos. Environ. 28 (1), 53–68.

De Hoogh, K., Chen, J., Gulliver, J., et al., 2018. Spatial PM2.5, NO2, O3 and BC models for Western Europe-Evaluation of spatiotemporal stability. Environ. Int. 120, 81–92.

Di, Q., Dominici, F., Schwartz, J.D., 2017a. Air pollution and mortality in the medicare population. N. Engl. J. Med. 377 (15), 1498–1499.

Di, Q., Rowland, S., Koutrakis, P., Schwartz, J., 2017b. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. J. Air Waste Manag. Assoc. 67 (1), 39–52.

Dicker, D., Nguyen, G., Abate, D., et al., 2018. Global, regional, and national age-sex-specific mortality and life expectancy, 1950-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 392, 1684–1735, 10159.

Feng, H., Zou, B., Tang, Y., 2017. Scale-and region-dependence in landscape-PM 2.5 correlation: implications for urban planning. Rem. Sens. 9, 918.

Hoek, G., Hoogh, B.K.D., Vienneau, D., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578.

Huang, L., Zhang, C., Bi, J., 2017. Development of land use regression models for PM2.5, SO2, NO2 and O3 in Nanjing, China. Environ. Res. 158, 542–552.

Huang, X., Ding, A., Gao, J., et al., 2020. Enhanced secondary pollution offset reduction of primary emissions during COVID-19 lockdown in China. Natl. Sci. Rev 0, 1–9.

Lee, H., Myung, W., Jeong, B.H., et al., 2018. Short- and long-term exposure to ambient air pollution and circulating biomarkers of inflammation in non-smokers: a hospital-based cohort study in South Korea. Environ. Int. 119, 264–273.

Liu, R., Ma, Z., Liu, Y., et al., 2020. Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: a machine learning approach. Environ. Int. 142, 105823.

Ma, Z., Hu, X., Sayer, A.M., et al., 2016. Satellite-based spatiotemporal trends in PM2.5 concentrations: China, 2004-2013. Environ. Health Perspect. 124 (2), 184–192.

Ma, R., Ban, J., Wang, Q., Li, T., 2020. Statistical spatial-temporal modeling of ambient ozone exposure for environmental epidemiology studies: a review. Sci. Total Environ. 701, 134463.

Ministry of Ecology and Environment of the People's Republic of China, 2019. Ecological and Environmental Bulletin in 2018. http://www.mee.gov.cn/hjzl/sthjzk/zghjzkgb/.

National Bureau of Statistics, 2020. Tabulation on the Population Census of the People's Republic of China by County. http://www.mee.gov.cn/hjzl/sthjzk/zghjzkgb/.

Peng, R.D., Samoli, E., Pham, L., et al., 2013. Acute effects of ambient ozone on mortality in Europe and North America: results from the APHENA study. Air Qual Atmos Health 6 (2), 445–453.

Ren, X., Mi, Z., Georgopoulos, P.G., 2020. Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: modeling ozone concentrations across the contiguous United States. Environ. Int. 142, 105827.

Son, Y., Osornio-Vargas, A.R., O'Neill, M.S., et al., 2018. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. Sci. Total Environ. 639, 40–48.

Wang, M., Keller, J.P., Adar, S.D., et al., 2015. Development of long-term spatiotemporal models for ambient ozone in six metropolitan regions of the United States: the MESA air study. Atmos. Environ. 123 (A), 79–87.

Wang, M., Sampson, P.D., Hu, J., et al., 2016. Combining land-use regression and chemical transport modeling in a spatiotemporal geostatistical model for ozone and PM2.5. Environ. Sci. Technol. 50 (10), 5111–5118.

Weagle, C.L., Snider, G., Li, C., et al., 2018. Global sources of fine particulate matter: interpretation of PM2.5 chemical composition observed by SPARTAN using a global chemical transport model. Environ. Sci. Technol. 52 (20), 11670–11681.

Wolf, K., Cyrys, J., Harcinikova, T., et al., 2017. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. Sci. Total Environ. 579, 1531–1540.

Wong, C.M., Vichit-Vadakan, N., Kan, H., Qian, Z., 2008. Public Health and Air Pollution in Asia (PAPA): a multicity study of short-term effects of air pollution on mortality. Environ. Health Perspect. 116 (9), 1195–1202.

Yang, B.Y., Qian, Z.M., Vaughn, M.G., et al., 2017. Is prehypertension more strongly associated with long-term ambient air pollution exposure than hypertension? Findings from the 33 Communities Chinese Health Study. Environ. Pollut. 229, 696–704.

Yang, B.Y., Qian, Z.M., Li, S., et al., 2018. Ambient air pollution in relation to diabetes and glucose-homoeostasis markers in China: a cross-sectional study with findings from the 33 Communities Chinese Health Study. Lancet Planet Health 2 (2), e64–e73.

Yin, P., Chen, R., Wang, L., et al., 2017. Ambient ozone pollution and daily mortality: a nationwide study in 272 Chinese cities. Environ. Health Perspect. 125 (11), 117006.

Zeger, SL, Thomas, D, Dominici, F, et al., 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environmental Health Perspectives 108 (5), 419–426.

Zhan, Y., Luo, Y., Deng, X., et al., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut. 233, 464–473.

Zhao, C., Wang, Q., Ban, J., et al., 2020. Estimating the daily PM2.5 concentration in the Beijing-Tianjin-Hebei region using a random forest model with a $0.01^\circ \times 0.01^\circ$ spatial resolution. Environ. Int. 134, 105297.