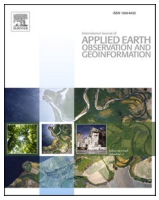
International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102955

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



Ground-level ozone estimation based on geo-intelligent machine learning by fusing in-situ observations, remote sensing data, and model simulation data

Jiajia Chen [a], Huanfeng Shen [a,b,*], Xinghua Li [c], Tongwen Li [d], Ying Wei [e]

[a] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China
[c] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[d] School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai 519082, China
[e] Institute of Urban Meteorology, China Meteorological Administration, Beijing 100089, China

ARTICLE INFO

Keywords:
Near-surface ozone estimation
Light gradient boosting machine
Spatio-temporal correlation
Ozone profile of model simulation
S5P-TROPOMI

ABSTRACT

In recent years, near-surface ozone (O_3) pollution has been increasing, seriously endangering both the ecological environment and human health. Accurately monitoring spatially continuous surface O_3 is still difficult with only remote sensing observations. In this paper, to address this issue, we propose a method for estimating surface O_3 by fusing multi-source data, including in-situ observations, O_3 precursors obtained by remote sensing, and model simulation data, including O_3 profile data and reanalysis products of meteorological and radiative elements. The estimation method is geo-intelligent light gradient boosting (Geoi-LGB) which takes into account both the spatial and temporal geographical correlation based on the standard LGB model. The spatio-temporal autocorrelation factors of the site observations are also constructed and added into the input variables. In a case study of China, centered on North China in 2019, the Geoi-LGB method obtained a root-mean-square error of 10.25 μg/m^3, a mean absolute error of 7.30 μg/m^3, and a coefficient of determination of 0.912 under the site-based cross-validation strategy. The proposed method has the advantages of being able to obtain a higher accuracy than some of the popular O_3 estimation models. Furthermore, the excellent spatial mapping ability of the Geoi-LGB method was demonstrated, in that about 85 % of the sites had an annual average absolute error of less than 10 μg/m^3. We believe that this study could provide some important reference information for the accurate estimation of ground-level O_3.

1. Introduction

Ground-level ozone (GLO) is a photochemical air pollutant. Alongside the rapid economic development and urbanization process, ozone (O_3) pollution is becoming more and more serious (Brauer et al., 2016; Lu et al., 2018). As a strong oxidizer, GLO can have negative effects on human health, including a low birth weight or premature delivery of infants (Wang et al., 2021a; Yang et al., 2020), an increase in the risk of anxiety or depression (Zhao et al., 2020), promotion of cardiovascular and respiratory diseases (Li et al., 2021b; Neidell and Kinney, 2010; Tian et al., 2020) and even premature death in ordinary people (Ito et al., 2005; Maji et al., 2019). In addition, in terms of ecosystem and climate change, O_3 affects the yields of many crops, thus causing national economic losses, and has an influence on the atmospheric temperature, due to its characteristic of absorbing ultraviolet radiation (Feng et al., 2019; Feng and Kobayashi, 2009; Li et al., 2021a; Manning and Tiedemann, 1995; Schauberger et al., 2019). Therefore, the accurate monitoring of GLO can help governments to identify the effects of O_3 hazards and undertake a more reasonable evaluation of the existing O_3 control measures.

At present, the main way of monitoring GLO concentration is the national air environment monitoring networks, which have the advantages of high precision and temporal continuity. However, this approach is limited to a discrete point distribution, and thus cannot meet the requirement for spatially continuous monitoring. The idea of point-surface fusion has also been widely used for the mass concentrations estimation of the air pollutant (Li et al., 2017c; Zhang et al., 2017). This process requires not only the station monitoring data, but also the

* Corresponding author at: School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China.
E-mail address: shenhf@whu.edu.cn (H. Shen).

https://doi.org/10.1016/j.jag.2022.102955
Received 11 May 2022; Received in revised form 16 July 2022; Accepted 1 August 2022
Available online 11 August 2022

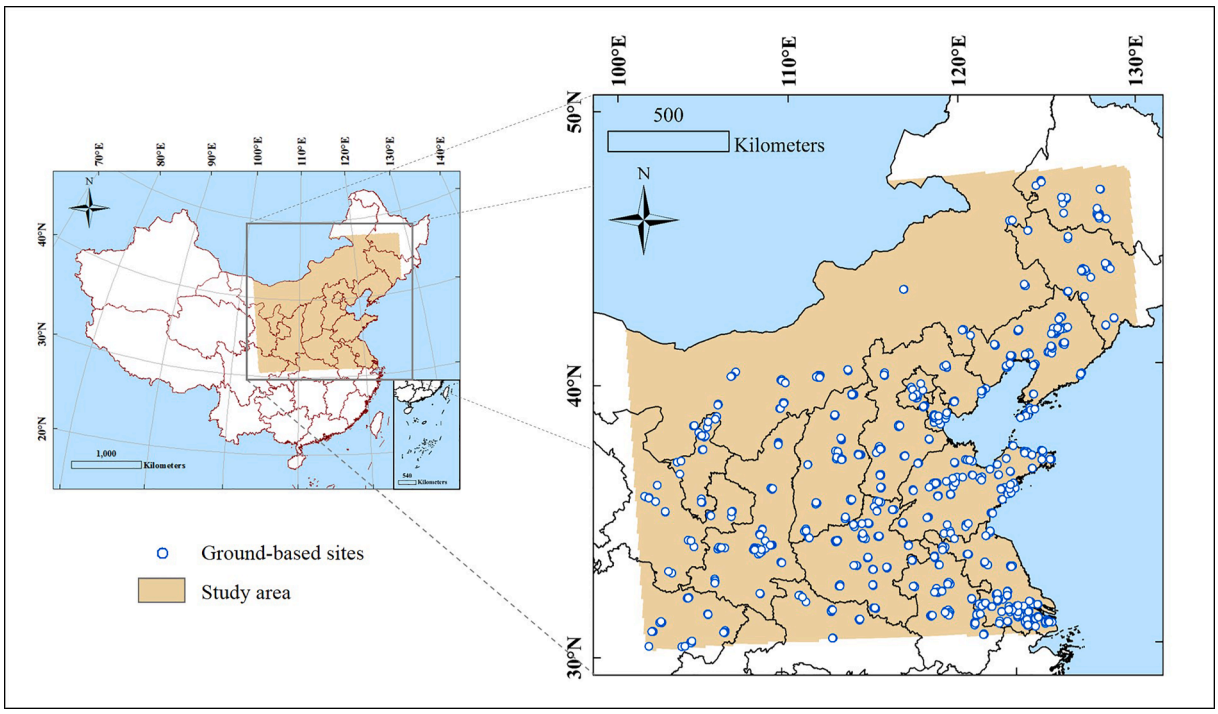


Fig. 1. The study area.

introduction of certain environmental factors related to the estimation components (Chameides et al., 1992). To date, for atmospheric O_3 related products, there are two main categories: model simulation and remote sensing, which are spatially continuous and have potential to be introduced into the point-surface fusion framework.

The common atmospheric environment simulation models, such as the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem), require both a ground emission inventory and a weather field as the input (Sicard et al., 2021; Wang et al., 2009; Zhang et al., 2010). The simulation results obtained by these models for the near-ground O_3 profile can be relatively close to the real situation, and the model accuracy has been improved in recent years (Travis and Jacob, 2019; Xue et al., 2020). However, due to the inaccuracy and hysteresis property of emission inventories caused by the "bottom-up" acquisition strategy and the errors in the process of simulation, there can still be a certain gap between the model results and the monitoring values of ground stations.

Remote sensing based atmospheric O_3 products have been gradually used to estimate GLO, including total O_3 column concentrations (Wang et al., 2021b; Wang et al., 2022) and O_3 profile products (Liang et al., 2019; Wang et al., 2022). Although the global products of total O_3 column concentrations can be considered as mature (Peng et al., 2016), the near-surface O_3 concentration accounts for less than 10 % of its total column. Therefore, it is difficult to directly construct its relationship with GLO. Furthermore, the current research results have shown that the strength of satellite detection of O_3 in the lower atmosphere is weak (Chatfield and Esswein, 2012; Wang et al., 2019a), because of the difficulty of overcoming the interference of the O_3 layer. Hence, the direct use of total column concentration or O_3 profile products in modeling can result in a great deal of uncertainty and theoretical defects. Some scholars have used O_3 precursor data in the estimation of GLO (Chameides et al., 1992; Fu et al., 2007; Wang et al., 2017; Zhang et al., 2020). It is known that ambient O_3 is generally formed from nitrogen oxides and volatile organic compounds, under the exposure of direct sunlight (Yadav et al., 2016; Zhang et al., 2018). Therefore, as respective representatives, nitrogen dioxide (NO_2) and formaldehyde (HCHO) have been gradually used in the estimation of GLO. These precursor products come principally from the anthropogenic emissions at the near-surface, and have a certain chemical reaction correlation between themselves and GLO (Zhang et al., 2020). However, the relationship between the precursors and GLO is indirect. Besides, no further consideration has been given to the radiative properties of O_3. In summary, the use of more data sources could have great potential for GLO estimation.

From the model perspective, the modeling methods for GLO estimation related to remote sensing have developed rapidly in recent years. Among the different models, machine learning based models have made great progress (Chiwewe and Ditsela, 2016; Li et al., 2020c; Xue et al., 2019). Compared with the traditional methods, such as the linear mixed effects model (Zhang et al., 2020), land use regression model (Kerckhoffs et al., 2015; Ren et al., 2020), and geographically weighted regression model (Van Donkelaar et al., 2016; Wang et al., 2019b; Zhang et al., 2020), machine learning based models can mine the nonlinear relationship between the independent variables and dependent variable. As a result, a higher accuracy can be obtained. In terms of machine learning based methods, bagging tree and boosting tree methods, such as random forest (RF) (Li et al., 2019; Zhan et al., 2018) and the gradient boosting machine (GBM) (Wei et al., 2021), have been widely used. The RF model, as a typical decision tree method, has the characteristics of stability and high accuracy, but as the number of decision trees is large, the space and time required for training will also be large (Zhan et al., 2018). GBM is a kind of machine learning method which optimizes the learning process using an addition model. In GBM, the algorithm using tree-based learners is called gradient boosting decision tree (GBDT). The most common base learner in GBDT is CART, used for classification and regression. However, for the GBDT, higher data dimensions will increase the computational complexity of the algorithm (Jerome, 2001). Its improved version, the extreme gradient boosting (XGB) model, still has a shortcoming that the strategy of level-wise leads the trees split even at the nodes with small gain, which brings unnecessary cost (Chen and Guestrin, 2016). Under circumstance of large amounts of data or high feature dimensions, the efficiency and scalability of these models still have room for improvement (Liu et al., 2020). To reduce the number of features and data without compromising the accuracy, a more advanced version, the light gradient boosting machine (LGB), was proposed in

Table 1
Data sources of this study.

Data type		Data source	Factor description	Abbreviation
In-situ observations		The China National Environmental Monitoring Center (CNEMC) (http://www.cnemc.cn)	Daily mean of hourly observations	SOzone
Model simulation	Three-dimensional O_3	RMAPS_Chem V2.0	O_3 concentration of the profile's lowest layer	MOzone
	Meteorological reanalysis data	GEOS-FP (https://portal.nccs.nasa.gov/datashare/gmao/geos-fp/das/)	10-m specific humidity	SH
			2-m air temperature	AT
			Planetary boundary layer height	PBLH
			Surface incoming shortwave flux	SWGDN
			TOA net downward shortwave flux	SWTNT
			Surface absorbed longwave radiation	LWGAB
Remote sensing		Sentinel-5P TROPOMI (https://s5phub.copernicus.eu/dhus/#/home)	TROPOMI NO_2 tropospheric column	TN
			TROPOMI HCHO tropospheric vertical column	TH

2017 (Ke et al., 2017). However, although machine learning models have made some progress in the estimation of GLO, they generally only fit the numerical corresponding relationship between GLO and its influencing factors based on single point modeling.

In general, on the one hand, we believe that insufficient data are introduced in the current methods, and using only a small amount of relevant data is not enough to achieve a high accuracy. On the other hand, the current statistical models, including the machine learning based methods, are usually used directly for modeling, and the spatio-temporal autocorrelation of the target data is seldom considered. In this work, aiming at the above two problems, we consider the use of multi-source data, including in-situ observations, O_3 profile data from model simulation, O_3 precursor tropospheric column concentration data, and meteorological reanalysis data, especially radiative factors. To better fuse the multi-source data, a geo-intelligent framework introducing the spatial and temporal autocorrelation of surface O_3 is proposed. We believe that this work could provide some important reference information for GLO spatial estimation. Section 2 provides details of the study area and the data used. Section 3 introduces the data preprocessing and the proposed method. Section 4 presents the proposed model performance from two aspects: a comparison with other typical methods and further verification. Section 5 discusses the influence of each variable on the model. Section 6 provides a summary of the work.

2. Study area and data source

2.1. Study area

The research region is shown in Fig. 1. The central area is North China (including Beijing, Tianjin, Hebei province, Shanxi province, and the central part of the Inner Mongolia Autonomous Region), which have more heavy air pollution in China (Ge et al., 2018; Gong et al., 2020). Ningxia Hui Autonomous Region, Shandong province, Liaoning province, Henan province, Shaanxi province, Jiangsu province, and the city of Shanghai are also included in the study area. The area range is 101° E–129° E, 30° N–48° N.

2.2. Data sources

Table 1 lists the data for 2019 used in this study. There are three main types of data: in-situ observations, model simulation data including three-dimensional O_3 and meteorological reanalysis factors, and remote sensing data, which are described in detail below.

2.2.1. In-situ near-surface O_3 observations

The in-situ near-surface O_3 concentrations were obtained from the China National Environmental Monitoring Center. There were 811 ground stations of this study area in 2019. The CNEMC provides the hourly ground-based O_3, in accordance with the technical specifications in HJ 818–2018, which states that the mean relative error for air pollutant should not exceed 5 %. Besides, according to the Ambient Air Quality Standards (GB 3095–2012), the hourly average concentration of air pollutants could be released only when the sampling time is at least 45 min per hour. As the selected satellite—the Copernicus Sentinel-5 Precursor satellite—transits once a day, the target for our estimation was the daily O_3 concentration. Differing from the commonly used index—the maximum daily 8-h average (MDA8)—the daily mean of the hourly observations was selected to maintain consistency with the model simulation results of O_3. Furthermore, the daily scale concentrations of GLO were acquired only when the effective monitoring hours were greater than 18 h. It is worth mentioning that these data were considered as the ground-truth values for the proposed GLO estimation framework. This type of O_3 data obtained by station monitoring is referred to as "SOzone" for short in this paper.

2.2.2. RMAPS_Chem O_3 profile

We simulated the O_3 profile data which is three-dimensional by the use of the RMAPS_Chem V2.0 system, which is an operational forecasting system for air pollution in North China, on the strength of the online coupled WRF-Chem regional chemical transport model. This method possesses the ability to reduce the error of offline calculation due to the spatial and temporal differences, and allows the study of the meteorological chemical feedback effect. The horizontal resolution of the RMAPS_Chem V2.0 system is 9 km × 9 km. The meteorological background field is provided by RMAPS-Short Term, which is composed of a WRF-based regional numerical weather prediction system and WRF data assimilation system (Xie et al., 2019; Zhong et al., 2020). The emission inventory data adopts the MEIC inventory of 2016 created by Tsinghua University (https://meicmodel.org; Li et al., 2017a; Zheng et al., 2018). The surface land categories are defined using the remote sense observation of Moderate Resolution Imaging Spectroradiometer (2000–2010). For the chemistry simulation, the carbon bond mechanism version Z (CBMZ) (Zaveri and Peters, 1999) and MOSAIC using 4 sectional aerosol bins (Zaveri et al., 2008) are used as the gas-phase chemistry module and the aerosol module, respectively. The simulation results for the O_3 profile had a total of 29 layers in the vertical direction per hour. We selected the data of the lowest layer, which is from the ground to about 100 m in height, and is the most closely related to near-surface O_3. The average over 24 h was obtained at a daily scale, which is abbreviated as "MOzone".

2.2.3. GEOS-FP reanalysis data

The GEOS-FP atmospheric reanalysis data was derived from the data assimilation system developed by the National Oceanic and Atmospheric Administration National Centers for Environmental Prediction. The output products have a spatial resolution of 0.3125° longitude by 0.25° latitude. In our study, six variables closely related to the process of O3 formation were picked: 10-m specific humidity (SH) (He et al., 2017), 2-m air temperature (AT) (Li et al., 2020a), planetary boundary layer height (PBLH) (Ma et al., 2011), and radiation-related variables (Chan

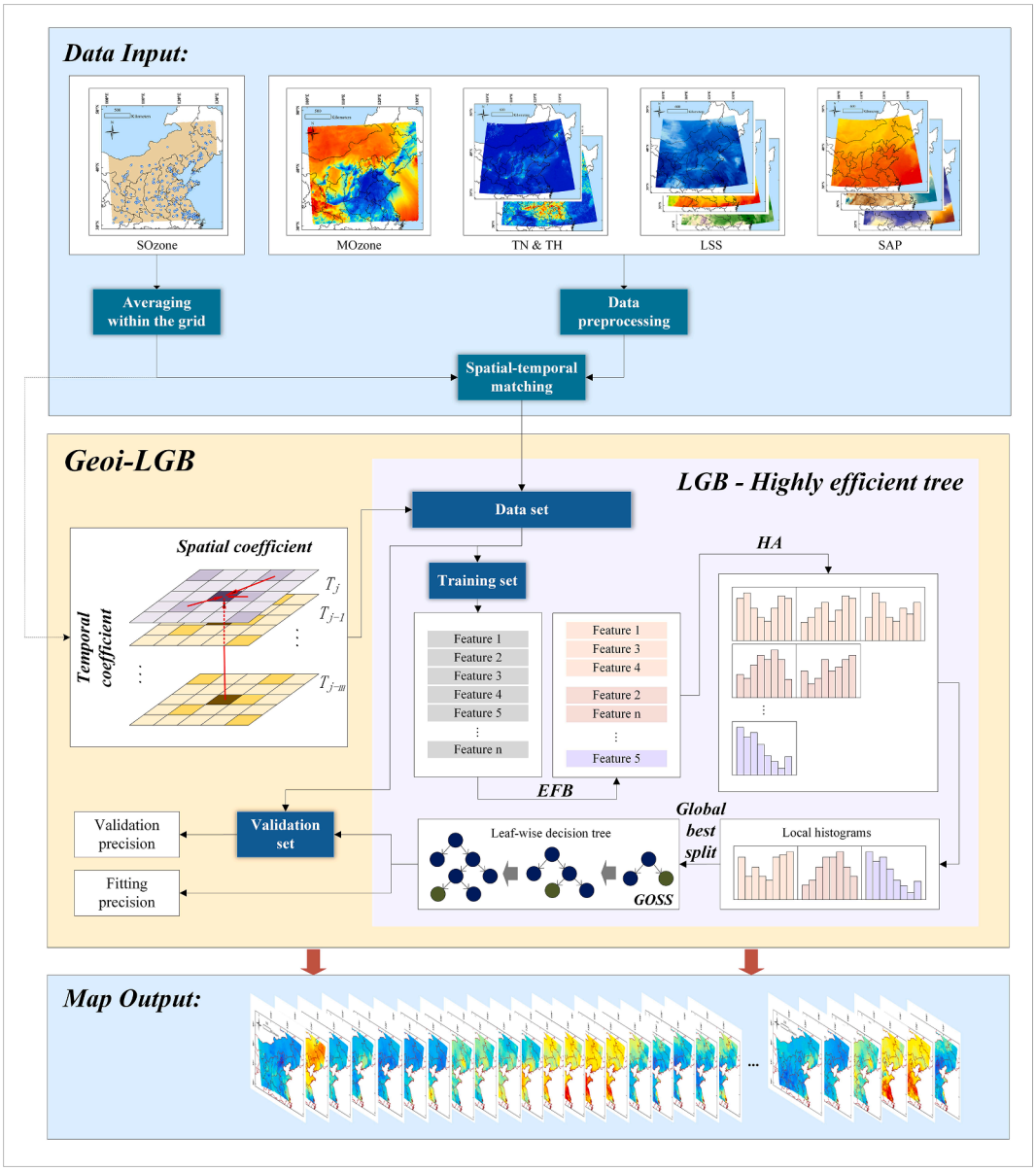


Fig. 2. Flowchart of the proposed method. (SAP includes SH, AT, and PBLH; LSS includes LWGAB, SWTNT, and SWGDN.).

and Chan, 2000) including surface incoming shortwave flux (SWGDN), TOA net downward shortwave flux (SWTNT), and surface absorbed longwave radiation (LWGAB).

2.2.4. TROPOMI tropospheric NO_2 and HCHO columns

The TROPOspheric Monitoring Instrument (TROPOMI) is carried on the Copernicus Sentinel-5 Precursor satellite, and has the highest spatial resolution among the atmospheric composition missions, such as MetOp, Aura, which is 3.5 km × 7.5 km at nadir. We selected the "TROPOMI NO_2 tropospheric column" and "TROPOMI HCHO tropospheric vertical column" products. An explanation about quality control can be found in Text S1 (Supported by Fig. S1, Table S1, and S2) of the Supplementary Materials. The TROPOMI NO_2 tropospheric column and TROPOMI HCHO tropospheric vertical column are respectively referred to as TN and TH in this paper.

Normalized difference vegetation index (NDVI) from the Moderate Resolution Imaging Spectroradiometer product—MOD13C1 (https://ladsweb.nascom.nasa.gov/) and digital elevation model (DEM) data from the Shuttle Radar Topography Mission (https://srtm.csi.cgiar.org/), were introduced into the proposed model initially, but it was found that the two kinds of data had some negative effects on the results under the framework proposed in this paper, which is elaborated in detail in Section 5.2.

2.3. Variables features

The descriptive statistics for the featured variables including the minimum, maximum, mean, and standard deviation are listed in Table S3 of the Supplementary Materials. It can be seen that the numerical distributions of the different elements are quite different. In addition, the linear correlation between each factor and SOzone is shown in Fig. S2 of the Supplementary Materials. As shown in Fig. S2, radiation data, MOzone, and meteorological data are strongly correlated with SOzone. Relatively speaking, the correlation with TH is poor. This shows that the traditional method making use of the linear assumption is relatively limited in modeling ability. Hopefully, with the ability of machine learning to mine nonlinear relationships, elements with weak linear relationships may play an important role.

3. Method

The objective of this study was to predict spatially continuous GLO concentrations from a series of environmental data, in the case of the data from discrete ground stations being considered as the ground-truth values. A complete flowchart for the proposed method is shown in Fig. 2. The data preprocessing and model building are described in detail below.

3.1. Data preprocessing

We constructed grids with a spatial resolution of 0.10°in the study area. The ground-level O_3 data after quality control observed from the monitoring sites in each grid cell were averaged. We reprojected all the variables used in this study to the consistent projection coordinate system. The simulated O_3 extracted from the O_3 profile, satellite precursors, and environmental factors were all resampled to the grid. The resampling method was the nearest neighbor interpolation on the grid scale. A complete database was formed after the data integration and spatiotemporal matching. The "Data Input" part in Fig. 2 depicts this process. With the accomplishment of the data preprocessing, the total number of data records was about 170,000.

3.2. Proposed model

3.2.1. Light gradient boosting

Some previous algorithm comparison studies found that the GBM is more competitive in the estimation methods of atmospheric compositions based on machine learning, even comparable to deep learning, and has a faster calculation speed (Li et al., 2020b; Wang et al., 2021b; Wei et al., 2021). LGB is an evolved version of the GBM model provided by Microsoft Research and released in 2017 (Ke et al., 2017). Compared with other algorithms in GBM category, the LGB model has many advantages, such as superior training efficiency, lower memory usage, and higher accuracy. The LGB model uses second-order approximation to minimize the objective function, which can quickly optimize the objective:

$$L^{(t)}=\sum_{i=1}^{n}\left[l\left(y_i,\hat{y}^{(t-1)}\right)+\partial_{y^{(t-1)}}l\left(y_i,\hat{y}^{(t-1)}\right)f_t(x_i)+\frac{1}{2}\partial^2_{y^{(t-1)}}l\left(y_i,\hat{y}^{(t-1)}\right)f_t^2(x_i)\right]+\Omega(f_t) \tag{1}$$

Where t represents the number of iterations; $L^{(t)}$ is the objective function of the t th iteration solution; n is the instance number of the input dataset; l is the loss function used to distinguish the difference between the predicted value $\hat{y}$ and the target value y_i of the i th instance at the t th iteration; ∂ and ∂^2 represent the first and second-order gradients of the loss function; $f_t(x)$ represents the corresponding increment; and $\Omega(f)$ is the regularization term.

The three most important improvements of the LGB model (depicted in the "LGB - Highly efficient tree" part of Fig. 2) can be summarized as: 1) Exclusive feature bundling (EFB); 2) Histogram-based algorithm (HA); 3) Gradient-based one-side sampling (GOSS). More detailed information can be found in the related paper (Ke et al., 2017).

The mapping relationship between the independent variables, i.e., MOzone, TN, TH, SH, AT, PBLH, LWGAB, SWGDN, and SWTNT and the dependent variable—SOzone—of the standard LGB can be summarized as shown in Eq. (2):

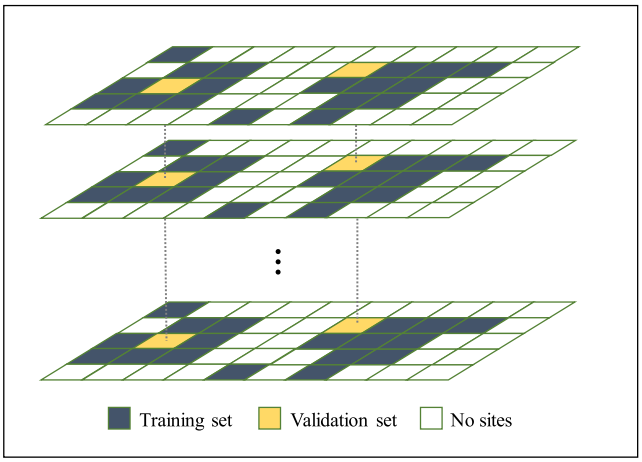


Fig. 3. Site-based cross-validation.

$$SOzone = LGB(MOzone, TN, TH, SH, AT, PBLH, LWGAB, SWGDN, SWTNT) \tag{2}$$

3.2.2. Geo-Intelligent light gradient boosting

Eq. (2) is based on single point modeling, without considering the spatiotemporal association between points. However, according to the first law of geography, all things or phenomena are connected in geography (Goodchild, 2009), including atmospheric components. To further improve the accuracy of the GLO estimation, this proposed method was inspired by its core idea—the connections between the concentrations of GLO in different spatial locations are geographically close to each other than those that are far away. Specifically, the spatial correlation between adjacent sites and the predicted location, and the temporal correlation between adjacent dates and the predicted date, were both incorporated in the model. This process is depicted in the diagram on the left side in the middle part of Fig. 2. The formulas for the spatial and temporal coefficients are described in Eq. (3) and Eq. (4), respectively, absorbing the more advanced geo-intelligent idea of fine particulate matter ($PM_{2.5}$) retrieval work (Shen et al., 2018). After adjusting the parameters of m and n according to experience, the results of the precision comparison show that the accuracy is the highest when the values are 3 and 5, respectively.

$$SC=\frac{\sum_{i=1}^{n}\varepsilon_{s,i}SOzone_i}{\sum_{i=1}^{n}\varepsilon_{s,i}},\ \varepsilon_{s,i}=\frac{1}{\Delta s_i^2} \tag{3}$$

$$TC=\frac{\sum_{j=1}^{m}\varepsilon_{t,j}SOzone_j}{\sum_{j=1}^{m}\varepsilon_{t,j}},\ \varepsilon_{t,j}=\frac{1}{\Delta t_j^2} \tag{4}$$

Where SC and TC refer to the spatial and temporal coefficients in the target grid cell of the target day, respectively; n refers to the number of nearest neighbor sites on that day; m refers to the number of the nearest day; $\varepsilon_{s,i}$ refers to the spatial weight of site i, which is the inverse square of the distance Δs_i to the target grid cell; $\varepsilon_{t,j}$ refers to the temporal weight of day j, which is the inverse square of the time distance Δt_j (calculated by day of year) to the target day; $SOzone_i$ refers to the O_3 mass concentration of site i; and $SOzone_j$ refers to the O_3 mass concentration of day j.

Based on the standard LGB model, the geo-intelligent LGB (Geoi-LGB) framework with spatio-temporal correlation elements—SC and TC, can be rewritten as Eq. (5):

$$SOzone = Geoi\text{-}LGB(MOzone, TN, TH, SH, AT, PBLH, LWGAB, SWGDN, SWTNT, SC, TC) \tag{5}$$

Table 2
Validation accuracies of each model.

Model	RMSE (μg/m^3)	R^2	MAE (μg/m^3)
DBN	17.81	0.735	13.60
RF	17.16	0.755	13.04
XGB	16.23	0.780	12.32
LGB	15.95	0.787	12.07
Geoi-LGB	10.25	0.912	7.30

According to the idea of tenfold cross-validation, we divided the whole dataset in a nine-to-one ratio into the training set and the validation set at every training time (Rodriguez et al., 2010). The training set was used for the construction of mapping relationship, and the validation set was applicable to evaluate the precision of the trained model. It should be noted that *SC* and *TC* in the training set and validation set were obtained separately, in order to avoid the information about the sites to be estimated involving in the training process. Based on the model with the optimal performance and the relevant raster data,

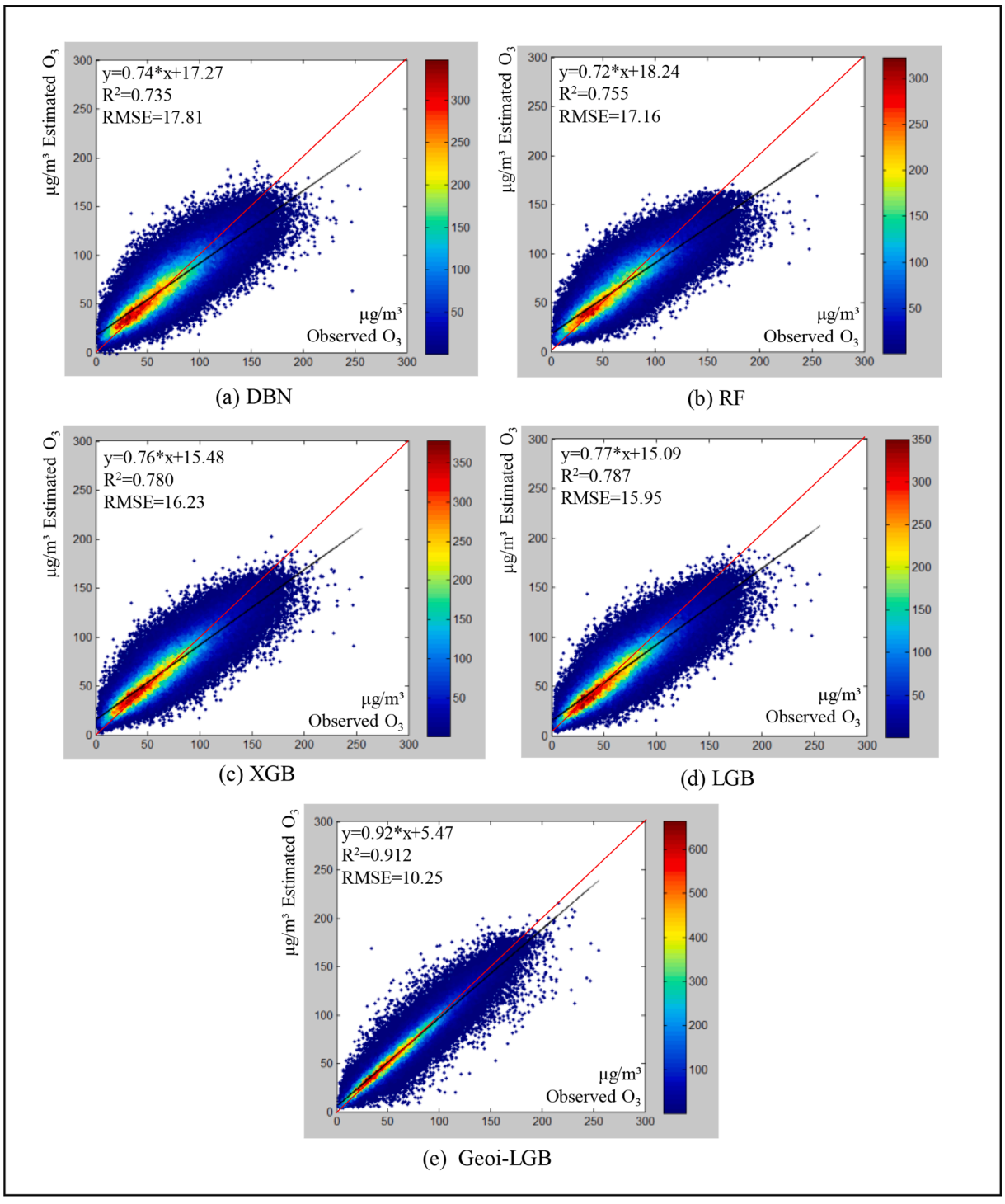


Fig. 4. Cross-validation scatter plots for 2019 obtained using the different models: (a) DBN; (b) RF; (c) XGB; (d) LGB; (e) Geoi-LGB.

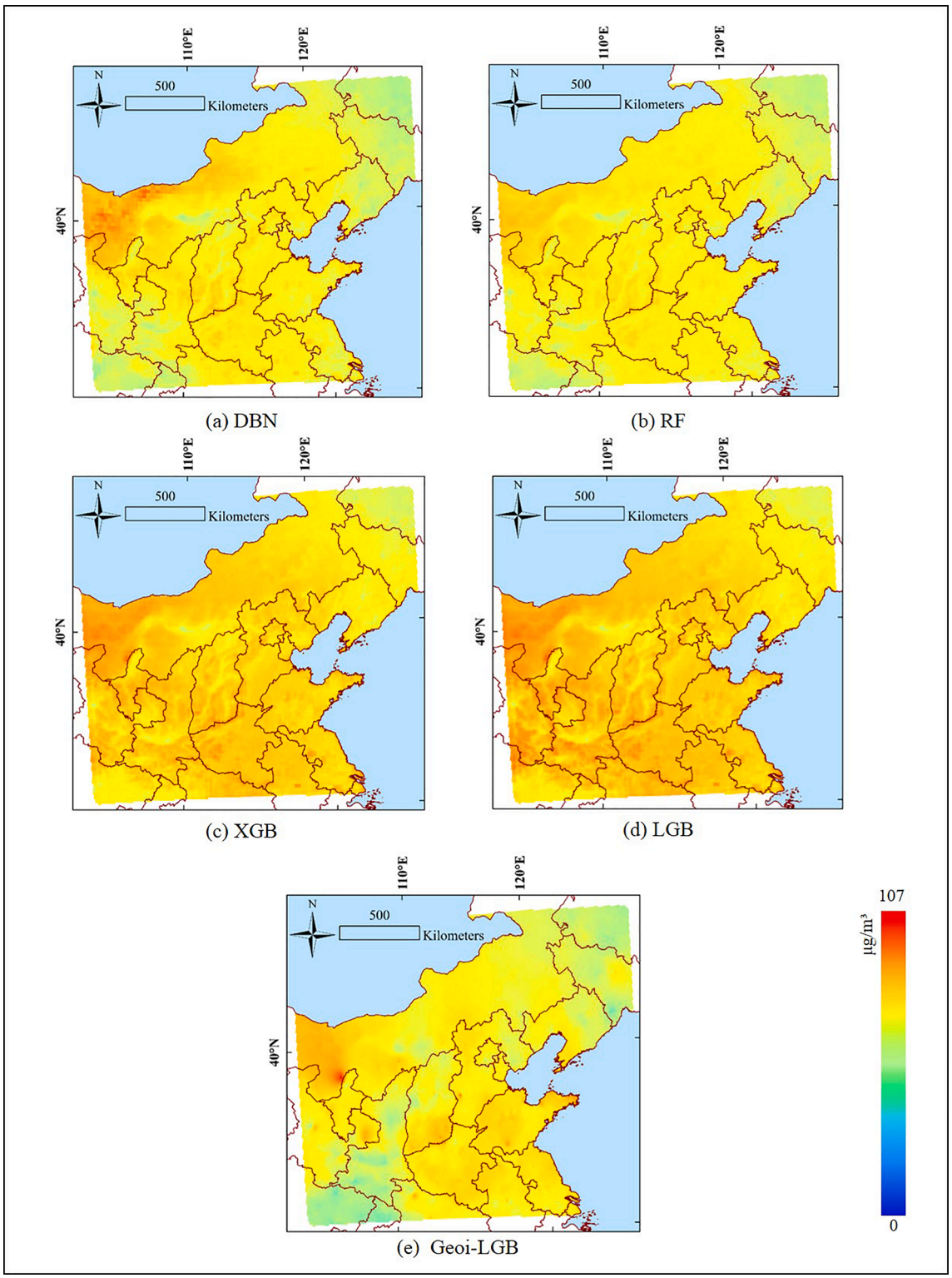


Fig. 5. Maps of the GLO concentration distribution for 2019 obtained using the different models: (a) DBN; (b) RF; (c) XGB; (d) LGB; (e) Geoi-LGB.

the spatially continuous GLO estimation results could be predicted, as depicted in the "Map Output" part in Fig. 2.

4. Experimental results

In this part, the difference between the Geoi-LGB model and the comparison methods in accuracy and mapping results is summarized, and the spatial and temporal prediction ability of the Geoi-LGB model is also analyzed. In order to show the spatial prediction ability of the model more directly, we chose the site-based validation approach (Fig. 3) rather than the common sample-based cross-validation approach. To evaluate the model precision, the coefficient of determination (R^2), the mean absolute error (MAE, $\mu g/m^3$), and the root-mean-square error (RMSE, $\mu g/m^3$) are used here as the validation metrics.

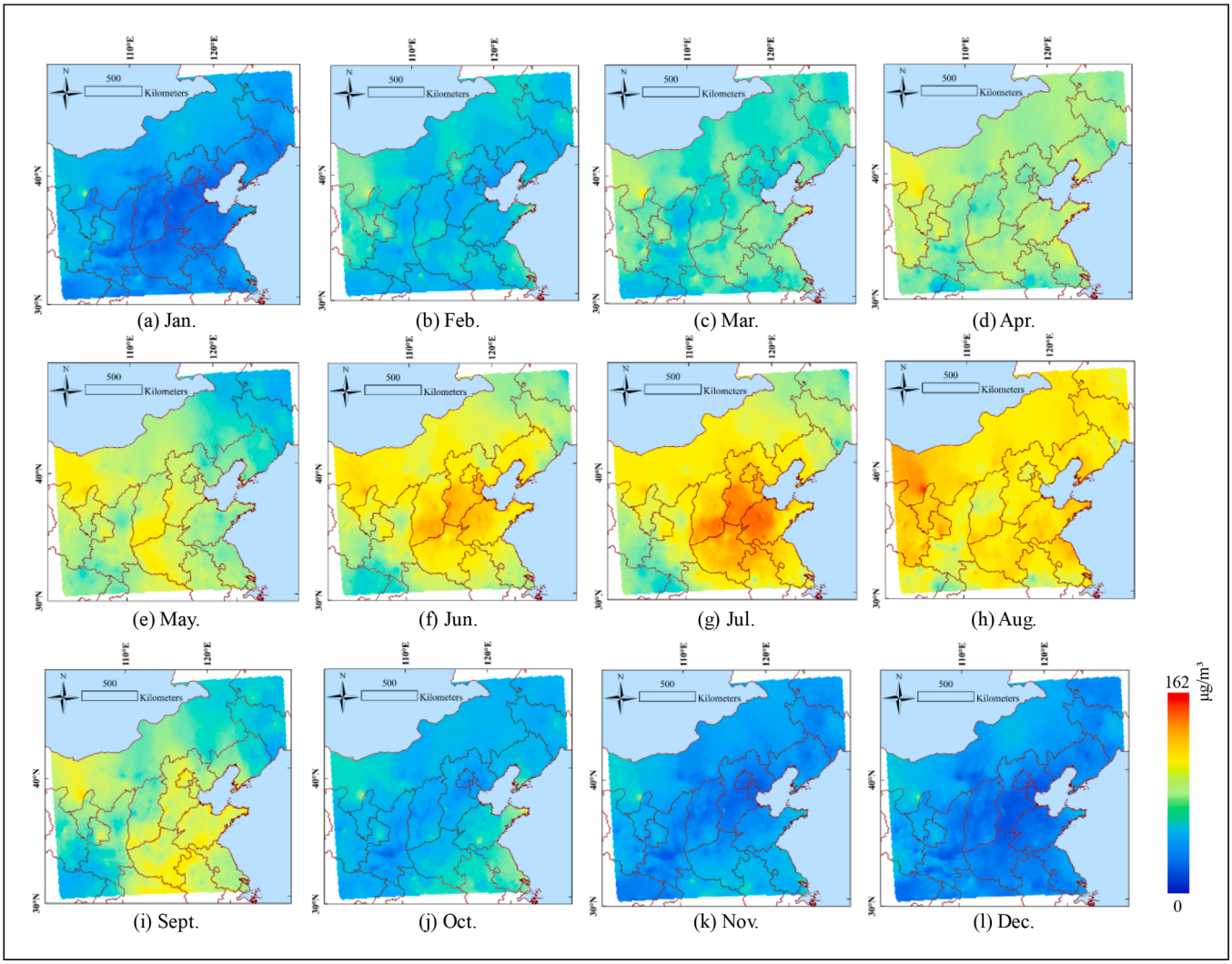


Fig. 6. The GLO estimation maps for each month of 2019.

4.1. Comparison with the state-of-the-art methods

To demonstrate the performance of the proposed model, we chose deep belief network (DBN) (Li et al., 2017b; Shen et al., 2020), RF (Li et al., 2019; Zhan et al., 2018), and XGB (Hu et al., 2022; Li et al., 2020b; Liu et al., 2020) models as the comparison machine learning approaches, which are frequently used for atmospheric composition ground concentration estimation (such as $PM_{2.5}$, O_3, NO_2, and sulfur dioxide). Detailed descriptions of these models can be found in the related papers. The training accuracies corresponding to each model is shown in Table S4 of the Supplementary Materials. The validation accuracy of each model after optimized parameter adjustment is shown in Table 2. The information of the hyper parameters setups of each model, the versions of corresponding libraries, and machine performance is presented in Table S5 of the Supplementary Materials. Without incorporating the geographical spatio-temporal correlation, the accuracy of the four models can be ranked as DBN < RF < XGB < LGB. It can be seen that the LGB model achieves the best performance, with the RMSE, R^2, and MAE values reaching 15.95 μg/m^3, 0.787, and 12.07 μg/m^3, respectively. Moreover, the XGB model is slightly inferior to the LGB model, although the results are relatively close. However, the LGB model performs clearly better than the RF and DBN models. Unexpectedly, the DBN model with the more complex model structure has the longest computation time and the lowest accuracy. The prediction ability of DBN is not good enough compared with the estimation works of other elements such as $PM_{2.5}$. This may have been caused by the combination of data inputs used in this work.

After incorporating the spatio-temporal correlation into the model, it is clear that the prediction accuracy of the Geoi-LGB model is significantly improved, with the RMSE, R^2, and MAE values reaching 10.25 μg/m^3, 0.912, and 7.30 μg/m^3 respectively. Compared with the LGB model, the RMSE and MAE are decreased by 5.70 μg/m^3 and 4.77 μg/m^3, and the R^2 is increased by 0.125. This confirms that near-surface O_3 has a strong correlation in time and space, which can be used to greatly improve the quantitative evaluation accuracy of GLO estimation.

Combining the cross-validation scatter plots of the five models for 2019 in Fig. 4 with the corresponding estimation maps for the annual mean GLO concentration in Fig. 5, the following conclusions can be drawn. Firstly, the spatial distributions of the GLO concentration values of the DBN, RF, and Geoi-LGB models are similar. Meanwhile, the results of the XGB and LGB models are clearly higher, overall. What is more, the DBN model has a different prediction effect, compared with the XGB and LGB models: while the overall values are generally similar to those of the RF and Geoi-LGB models, the predicted values in the northwest are clearly higher. Lastly, the effect of Geoi-LGB mapping is significantly affected by the spatio-temporal correlation of the ground observations.

The above cartographic differences may have been caused by the following reasons. Firstly, the DBN model has a more complex neural network structure. In the training process, the DBN model performs better on information of high and low values with a small sample size.

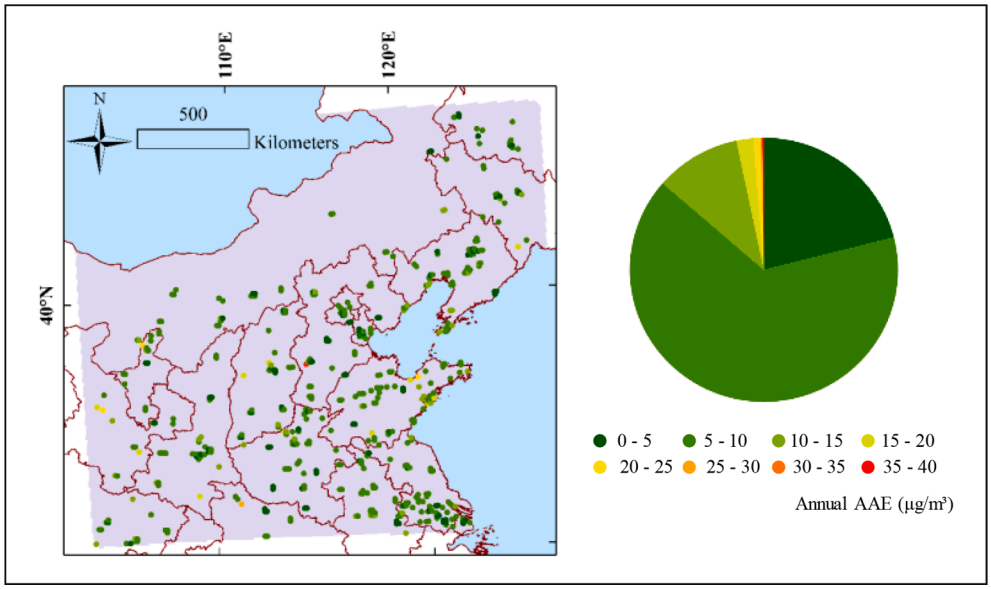


Fig. 7. The spatial distribution of the annual AAE obtained with the Geoi-LGB model under the site-based validation method.

However, the estimation and mining ability for such extreme values is not completely accurate, and although the slope of the estimated/observed values is closer to 1 than that of the RF model, the final accuracy is not high enough. The RF model shows a stable modeling capability, which is consistent with the good results obtained in other regression applications. This shows that bagging decision tree learning is relatively stable in terms of the prediction accuracy, and benefits from its integrated decision-making ability. The boosting tree, due to its mechanism of sequential minimization of the loss function, can effectively improve the prediction accuracy, but the mapping ability of the XGB model in this work is limited, which may be due to its approximation algorithm. Although the LGB model obtains the highest accuracy in the quantitative evaluation comparison, the predicted values of the LGB model are significantly higher, overall, and the degree of overestimation is the most serious. This means that, even though its optimization scheme is efficient, the simplified and acceleration algorithm also brings some problems, such as the histogram-based algorithm and the random selection of high gradient samples, as the price of efficiency. Clearly, the proposed Geoi-LGB model effectively corrects the high prediction results of the LGB model, and the incorporation of *SC* and *TC* has a greater impact on the mapping results than the other related factors.

In conclusion, the LGB model has outstanding advantages in terms of the quantitative evaluation, and the Geoi-LGB model, which considers the spatio-temporal correlation, greatly improves the overestimation of the LGB model. Overall, the proposed Geoi-LGB model shows an excellent performance in the quantitative evaluation, operational efficiency, and mapping effect.

4.2. Spatial and temporal analysis of the Geoi-LGB model estimation results

In this part, we analyze the predictive ability of the Geoi-LGB model in different space and time, respectively. Fig. 6 shows the estimated GLO maps for every month in 2019, where the monthly variation of GLO can be seen intuitively. The GLO concentration is low in winter (January, February, and December) and high in summer (from June to August), which is consistent with the light and radiation intensity. The cartographic effect also accords with the diffusion law of atmospheric circulation. Due to the consideration of the temporal and spatial autocorrelation, some local measurement values of individual monitoring stations are also clearly reflected.

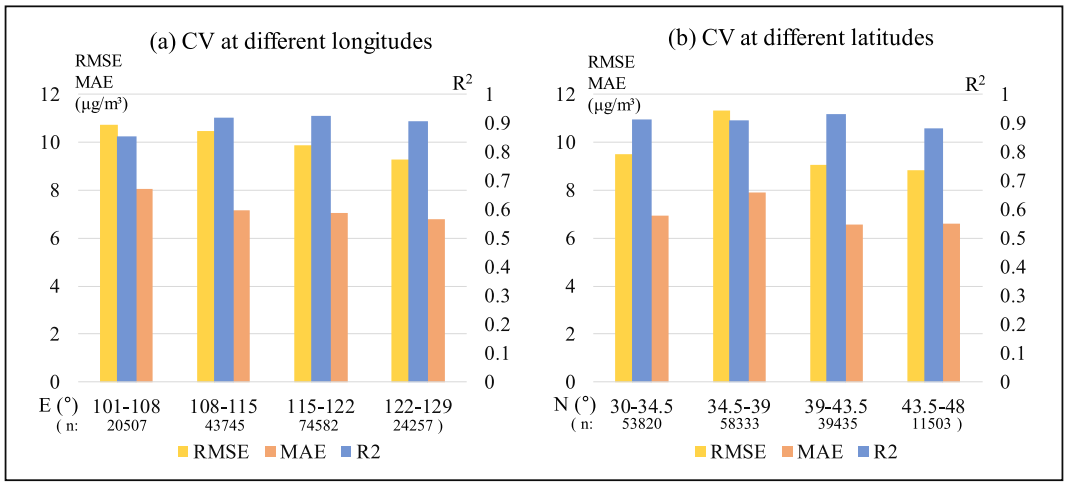


Fig. 8. Cross-validation comparison in different (a) longitudes and (b) latitudes.

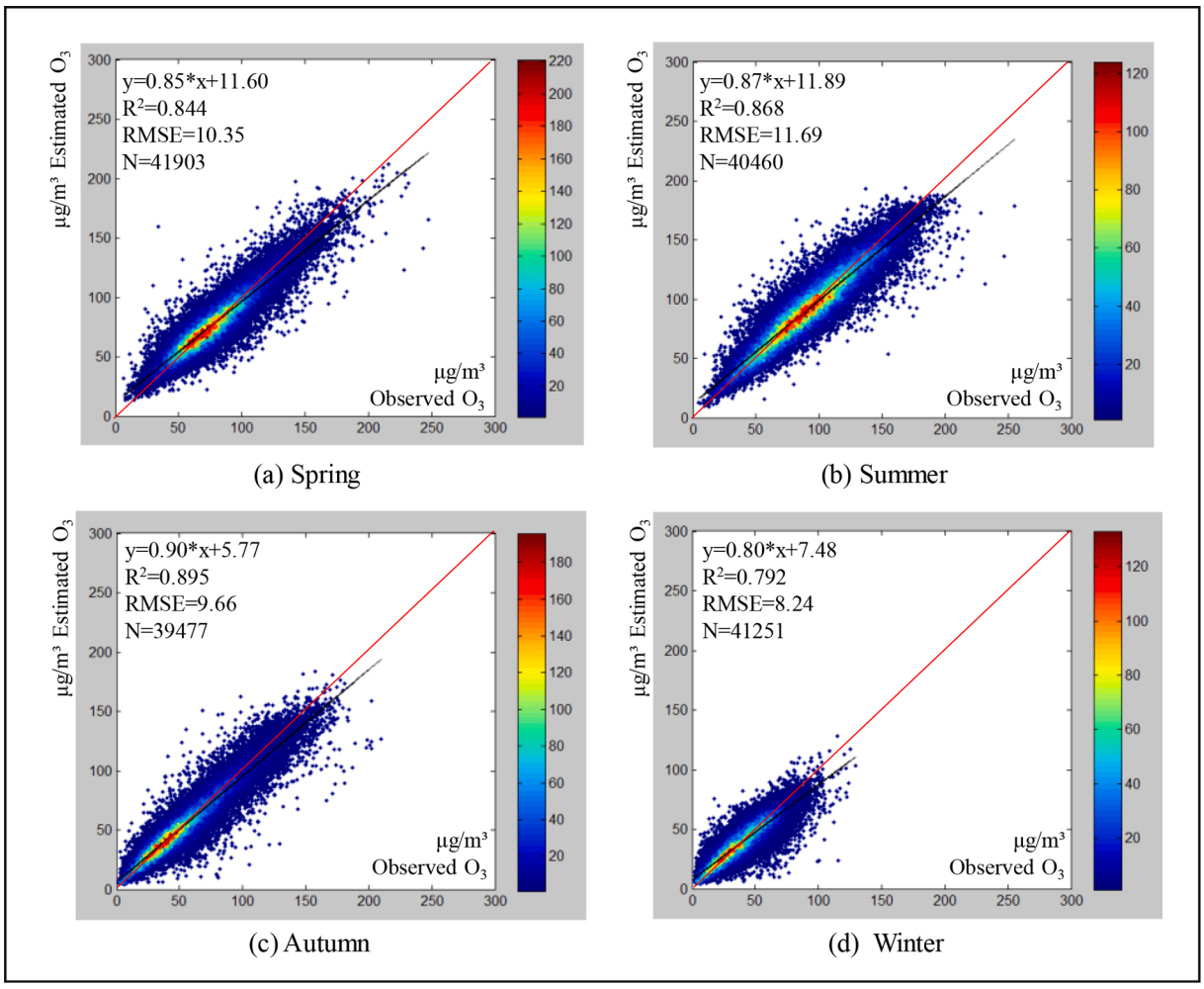


Fig. 9. Cross-validation comparison in the different seasons: (a) spring; (b) summer; (c) autumn; (d) winter.

4.2.1. Spatial difference of the predictive performance

We calculated the annual average absolute error (AAE) for each site using the Geoi-LGB model. The results are shown in Fig. 7, where it is obvious that the distribution of the annual AAE is relatively uniform. Among the results, those with an annual AAE of less than 5 μg/m^3 account for about 23 %, and there are only individual sites with slight discrepancies. In addition, about 85 % of the sites have an annual AAE of less than 10 μg/m^3, for which the spatial distribution is uniform and dispersed. Due to the advantage of the uniform distribution of the monitoring stations, the spatial distribution of the annual AAE is relatively homogeneous. The annual AAE values of more than 10 μg/m^3 may be due to the comprehensive effects of the other environmental data. Notably, the annual AAE in Beijing, Tianjin, and the Yangtze River Delta region with the densest monitoring stations is low, indicating that the

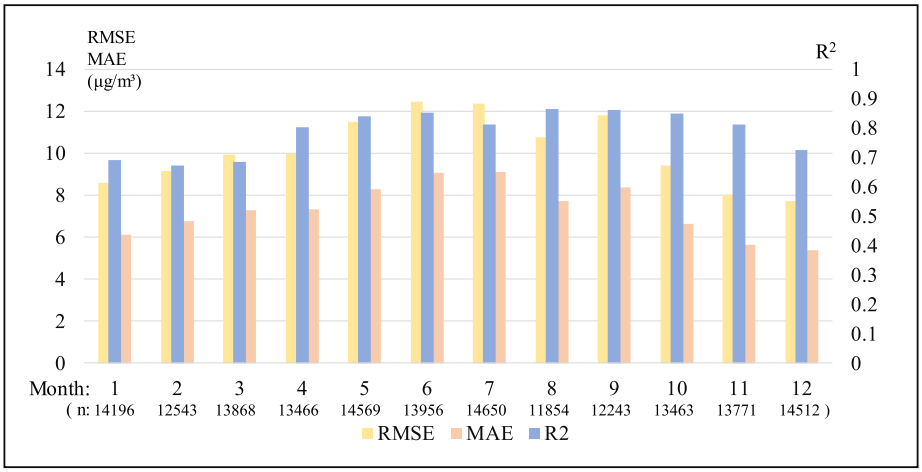


Fig. 10. Cross-validation comparison in the different months.

Table 3
Comparison of the validation accuracies after removing each individual factor.

Data	RMSE (μg/m^3)	R^2	MAE (μg/m^3)
All data	**15.95**	**0.787**	**12.07**
Without MOzone	**16.63**	**0.769**	**12.55**
Without TH	16.12	0.783	12.20
Without TN	16.25	0.780	12.34
Without TH, TN	**16.45**	**0.774**	**12.51**
Without SH	16.23	0.780	12.30
Without AT	16.99	0.762	12.76
Without PBLH	16.64	0.769	12.67
Without SAP	**18.63**	**0.710**	**14.17**
Without LWGAB	16.50	0.773	12.49
Without SWGDN	16.81	0.764	12.70
Without SWTNT	17.04	0.758	12.90
Without LSS	**18.73**	**0.707**	**14.33**

(SAP includes SH, AT, and PBLH; LSS includes LWGAB, SWTNT, and SWGDN.).

point density of sites is of great help to reduce the spatial prediction error.

Fig. 8 shows the model accuracy in different latitudes and longitudes. Fig. 8a indicates that the number of stations between 115 and 122 E° is relatively high and the accuracy is also relatively high. It is worth noting that, by comparing the western region (101–108° E) and part of the northeastern region (122–129° E), it can be found that the accuracy for the northeastern region is significantly higher than that for the western region, despite the numbers of samples being almost the same. This is likely because the western region has a larger latitude span, while the northeastern region is smaller and has a denser distribution of sites. Fig. 8b shows the situation of the region with a latitude of 43.5–48° N being the same as the region of 122–129° E, which are more concentrated in northeast China. Therefore, even though there are fewer stations, they are relatively dense, and a higher accuracy is obtained. Surprisingly, among the other three latitude ranges, the range of 34.5–39° N, which has the largest sample size, obtains the lowest accuracy. As can be seen in Fig. 1, this is because that the region starts from Shaanxi province in the west and ends in Shandong province in the east. Compared with the two adjacent latitude bands in the north and south, the longitude span of this region is larger. Therefore, the distribution of GLO is significantly different, which makes it more difficult to obtain an accurate prediction with the model.

4.2.2. *Temporal difference of the predictive performance*

To clarify the temporal difference of the performance of the Geoi-LGB model, Figs. 9-10 shows the model accuracies of different seasons and months. First of all, Fig. 9 compares the cross-validation results obtained in the spring (March–May), summer, autumn (September–November), and winter of 2019. In terms of the RMSE (the lower the better), the results are winter < autumn < spring < summer. In terms of the R^2 (the higher the better), the results are autumn > summer > spring > winter. The RMSE in summer is 3.45 μg/m^3 higher than that in winter, and the R^2 in winter is 0.103 lower than that in autumn. The reason for this inconsistency is that the RMSE is greatly affected by the observed values. Specifically, the order of the mean values of GLO concentration in each season—summer (92.97 μg/m^3) > spring (78.58 μg/m^3) > autumn (54.03 μg/m^3) > winter (37.48 μg/m^3), is numerically accordant with the obtained RMSE. Fig. 10 further shows the prediction accuracy of the model in different months. It can be seen that the RMSE and MAE have a law of first increasing and then decreasing in the monthly variation, which is consistent with the change of GLO concentration. The R^2 maintains a high value from April to November, indicating that the model has good predictive power in these months.

5. Discussion

For purpose of evaluating the effect of the various feature sources on the model, modeling, precision evaluation, and mapping were all carried out with the LGB model on the datasets after removing one variable at a time. The improvement brought by each data source could then be clearly demonstrated. In the following, the influence of the atmospheric components and meteorological reanalysis factors on the prediction results is analyzed. Furthermore, the inapplicability of surface elements such as NDVI and DEM is also discussed.

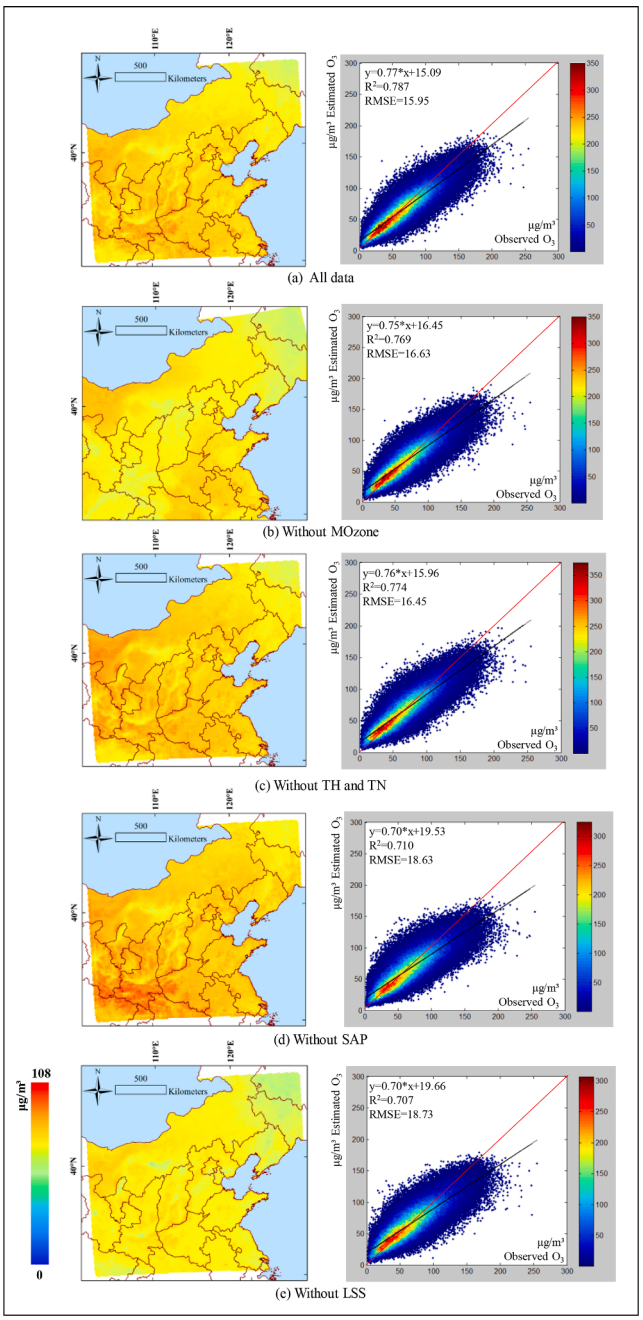


Fig. 11. Maps of the GLO concentration distribution and cross-validation scatter plots for 2019 with different data inputs: (a) all data; (b) without MOzone; (c) without TN and TH; (d) without SAP; (e) without LSS.

5.1. *The influence of each variable on the prediction results*

Firstly, the roles of the three atmospheric components, i.e., MOzone, TH, and TN, in the estimation of surface O_3 are evaluated. Rows 3–6 of Table 3 show the validation results of the corresponding model when MOzone, TH, and TN were respectively removed from the modeling. It can be seen that the RMSE and MAE decrease by 0.68 and 0.48 μg/m^3,

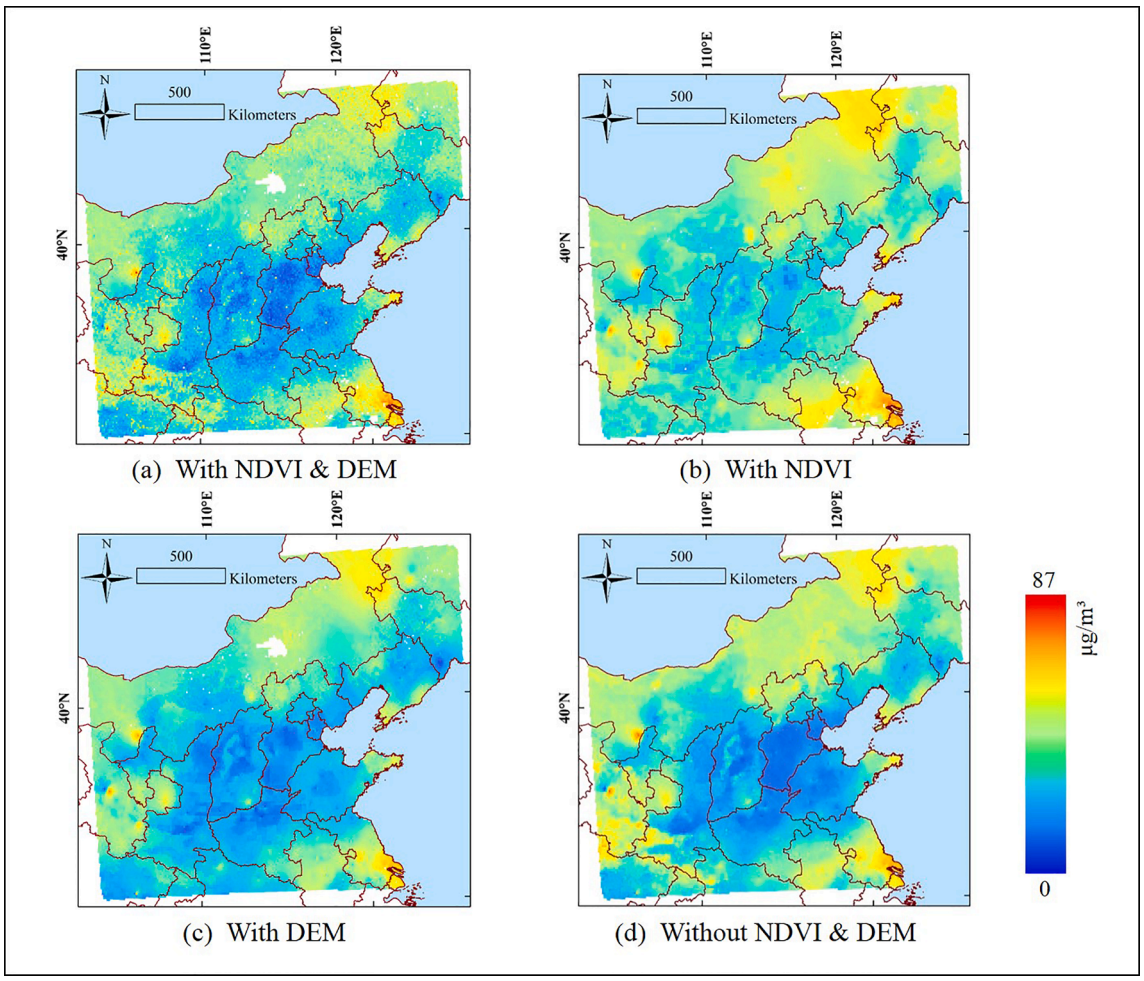


Fig. 12. Estimation maps for GLO on Jan. 1, 2019, in different circumstances: (a) with NDVI and DEM; (b) with NDVI; (c) with DEM; (d) without NDVI and DEM.

respectively, and the R^2 is increased by 0.018, after MOzone is included in the modeling, compared with the situation without MOzone. Similarly, after TH and TN are included in the modeling, the RMSE and MAE decrease by 0.50 and 0.44 μg/m^3, respectively, and the R^2 increases by 0.013. In general, the effects of the three kinds of atmospheric data in improving the model performance can be ranked as follows: MOzone > TN > TH, and the combined impact of TN and TH is slightly less than that of MOzone alone. These results indicate that both the TROPOMI remote sensing O_3 precursors and the model simulation O_3 have positive effects on improving estimation accuracy, and the latter has a more significant effect.

In order to show the influence of each variable on the mapping results, Fig. 11 shows the annual mean values of the prediction results. By comparing Fig. 11a and b, it is clear that, with MOzone participating in the modeling, the spatial prediction graph shows more spatial trends with fluidity. The spatial fluidity generated by the transport of atmospheric pollutants with meteorological elements is reflected in the mapping results. The reason for this is that one of the foundations of atmospheric model simulation results is the meteorological variable field. The slope of the scatter plot is also improved, and it can be seen that the low predicted values have been corrected. It is worth noting that the distribution of spatial traces brought by the introduction of MOzone is different from the estimated concentration of other atmospheric pollutants. The mapping results of Geoi-LGB in Fig. 5 more match with the common sense. It can be considered that the introduction of the spatio-temporal autocorrelation factors eliminates this unreasonable effect, and the TROPOMI remote sensing data are at important work in correcting the low values. From Fig. 11a and c, it can be concluded that TN and TH do not take an obvious effect in the spatial characterization, but they do contribute to correcting the low predicted values. Overall, the effect of MOzone is greater than that of the remote sensing precursor data.

In this study, six key variables were selected from the meteorological reanalysis data, according to the strong linear correlation with SOzone. From the quantitative evaluation results shown in rows 7–14 of Table 3, it can be seen that the influence of the six variables on the prediction accuracy can be ranked from high to low as follows: SWTNT > AT > SWGDN > PBLH > LWGAB > SH. The six variables can be divided into two categories: (1) conventional meteorological parameters (SH, AT, and PBLH, summarized by SAP); and (2) atmospheric radiation parameters (LWGAB, SWGDN, and SWTNT, summarized by LSS). In the following, the functions of these two types of data are analyzed. With the addition of SAP, the RMSE and MAE of the model are reduced by 2.68 μg/m^3 and 2.10 μg/m^3, respectively, and the R^2 is increased by 0.077. The corresponding changes of RMSE, MAE, and R^2 with the addition of LSS are 2.78 μg/m^3, 2.26 μg/m^3, and 0.080, respectively. Overall, LSS plays a slightly greater part in the results than SAP.

By comparing the mapping effects of Fig. 11a, d, and e, it is clear that the GLO estimation results obtained without considering LSS are generally low, with less spatial details. The spatial prediction results without SAP are significantly higher in the south of Shaanxi province, and there is no distinct difference in other areas. On the whole, LSS has a greater influence on the results than SAP in the aspect of mapping, in keeping with the quantitative evaluation. This also demonstrates the close relationship between solar radiation and GLO.

5.2. The inapplicability of NDVI and DEM to the proposed model

In our experiments in introducing related variables, it was found that the GLO estimation accuracy could be slightly improved after adding NDVI and DEM into the Geoi-LGB model. However, there are obvious problems in the visual effect, as shown in Fig. 12, which shows the estimation maps for GLO on Jan. 1, 2019, in different circumstances, i.e., with NDVI and DEM, with NDVI, with DEM, and without NDVI and DEM. It can be seen that when NDVI and DEM are combined into the model, the prediction map shows obvious spot-like noise traces. When only NDVI or DEM is removed, most of the daily-scale mapping results show obvious block traces. This may be because DEM and NDVI have relatively large spatial gradients in values, while the LGB model selects features with large information gain for the regression calculation, in order to reduce the calculation cost. Such a calculation mechanism amplifies the features with a large numerical variation, to a certain extent. It is also shown that the spatial estimation ability of the LGB method is affected by the numerical continuity of the correlated characters in space, to some extent.

6. Conclusion

In this paper, we have proposed a geo-intelligent highly efficient tree model—the Geoi-LGB—by taking into account the spatial and temporal geographical correlations to estimate the concentration of GLO, and validated the results with data from 2019. The Geoi-LGB model obtained an RMSE of 10.25 μg/m^3, an R^2 of 0.912, and a MAE of 7.03 μg/m^3, which represents high-precision estimation of GLO for 2019. When compared with the DBN, RF, and XGB, models, it was found that the LGB model has the advantage of a higher accuracy. In addition, the excellent spatial estimation ability of the Geoi-LGB model was also proved, in that about 85 % of the sites had an annual AAE of less than 10 μg/m^3. In terms of the seasonal difference, we found that the prediction ability in autumn is better than that in the other seasons, followed by spring and summer. The results showed that there is good correlation between the low-level data of the O_3 profile simulated by the model and the O_3 concentration at ground stations, which plays a significant part in advancing the model's accuracy. The introduction of two TROPOMI O_3 precursors and meteorological and radiation data from the GEOS-FP reanalysis data also improved the accuracy of the GLO estimation. Our work realizes the cascade of O_3 element in model simulation and machine learning method modeling, and mines the role of the output data of the former in the estimation of air pollutants in a statistical category method. Geo-intelligent consideration greatly improves our model accuracy. The method proposed in this paper is also applicable in other regions where relevant data are available. Although the current estimation process still has some room for improvement, such as the optimization algorithm used for the local heterogeneity and the quality control of remote sensing products, we believe that this study will provide some important reference information for the accurate estimation of GLO.

CRediT authorship contribution statement

Jiajia Chen: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Huanfeng Shen:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Xinghua Li:** Project administration, Supervision, Writing – review & editing. **Tongwen Li:** Formal analysis, Methodology, Writing – review & editing. **Ying Wei:** Data curation, 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.

Data availability

No data was used for the research described in the article.

Acknowledgments

We would like to express our sincere thanks to the China National Environmental Monitoring Center for the hard work of the in-situ near-surface O_3 observations, collection, management, and maintenances, the Copernicus Open Access Hub for providing the S5P-TROPOMI products, the Goddard Space Flight Center for developing the GEOS-FP reanalysis data, and other institutions for providing the relevant datasets. Most importantly, thanks for the foundation support of the National Key Research and Development Program of China (2019YFB2102900).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2022.102955.

References

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R.V., Dentener, F., Dingenen, R.v., Estep, K., Amini, H., Apte, J.S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J.L.T., Shaddick, G., Anderson, H.R., Vos, T., Forouzanfar, M.H., Burnett, R.T., Cohen, A., 2016. Ambient air pollution exposure estimation for the global burden of disease 2013. Environ. Sci. Technol. 50, 79.

Chameides, W.L., Fehsenfeld, F., Rodgers, M.O., Cardelino, C., Martinez, J., Parrish, D., Lonneman, W., Lawson, D.R., Rasmussen, R.A., Zimmerman, P., Greenberg, J., Mlddleton, P., Wang, T., 1992. Ozone precursor relationships in the ambient atmosphere. J. Geophys. Res.-Atmos. 97, 6037–6055.

Chan, C.Y., Chan, L.Y., 2000. Effect of meteorology and air pollutant transport on ozone episodes at a subtropical coastal Asian city, Hong Kong. J. Geophys. Res.-Atmos. 105, 20707–20724.

Chatfield, R.B., Esswein, R.F., 2012. Estimation of surface O3 from lower-troposphere partial-column information: Vertical correlations and covariances in ozonesonde profiles. Atmos. Environ. 61, 103–113.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: In ProceedIngs of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data MInIng, pp. 785–794.

Chiwewe, T.M., Ditsela, J., 2016. Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations. In: 2016 IEEE 14th International Conference on Industrial Informatics, pp. 58–63.

Feng, Z.Z., Kobayashi, K., 2009. Assessing the impacts of current and future concentrations of surface ozone on crop yield with meta-analysis. Atmos. Environ. 43, 1510–1519.

Feng, Z., Kobayashi, K., Li, P., Xu, Y., Tang, H., Guo, A., Paoletti, E., Calatayud, V., 2019. Impacts of current ozone pollution on wheat yield in China as estimated with observed ozone, meteorology and day of flowering. Atmos. Environ. 217, 116945.

Fu, T.-M., Jacob, D.J., Palmer, P.I., Chance, K., Wang, Y.X., Barletta, B., Blake, D.R., Stanton, J.C., Pilling, M.J., 2007. Space-based formaldehyde measurements as constraints on volatile organic compound emissions in east and South Asia and implications for ozone. J. Geophys. Res.-Atmos. 112, D06312.

Ge, B., Wang, Z., Lin, W., Xu, X., Li, J., Ji, D., Ma, Z., 2018. Air pollution over the North China Plain and its implication of regional transport: A new sight from the observed evidences. Environ. Pollut. 234, 29–38.

Gong, C., Liao, H., Zhang, L., Yue, X., Dang, R., Yang, Y., 2020. Persistent ozone pollution episodes in North China exacerbated by regional transport. Environ. Pollut. 265, 115056.

Goodchild, M.F., 2009. International Encyclopedia of Human Geography: First Law of Geography, 179–182. Elsevier Science.

He, J., Gong, S., Yu, Y., Yu, L., Wu, L., Mao, H., Song, C., Zhao, S., Liu, H., Li, X., Li, R., 2017. Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities. Environ. Pollut. 223, 484–496.

Hu, X., Zhang, J., Xue, W., Zhou, L., Che, Y., Han, T., 2022. Estimation of the Near-Surface Ozone Concentration with Full Spatiotemporal Coverage across the Beijing-

Tianjin-Hebei Region Based on Extreme Gradient Boosting Combined with a WRF-Chem Model. Atmosphere 13, 632.
Ito, K., De Leon, S.F., Lippmann, M., 2005. Associations between ozone and daily mortality: Analysis and meta-analysis. Epidemiology. 16, 446–457.
Jerome, H.F., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29, 1189–1232.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. NIPS 2017. Long Beach, CA, USA, 4-9 December 2017, 1-9.
Kerckhoffs, J., Wang, M., Meliefste, K., Malmqvist, E., Fischer, P., Janssen, N.A.H., Beelen, R., Hoek, G., 2015. A national fine spatial scale land-use regression model for ozone. Environ. Res. 140, 440–448.
Li, R., Cui, L., Meng, Y., Zhao, Y., Fu, H., 2019. Satellite-based prediction of daily SO2 exposure across China using a high-quality random forest-spatiotemporal Kriging (RF-STK) model for health risk assessment. Atmos. Environ. 208, 10–19.
Li, C., Gu, X., Wu, Z., Qin, T., Guo, L., Wang, T., Zhang, L., Jiang, G., 2021a. Assessing the effects of elevated ozone on physiology, growth, yield and quality of soybean in the past 40 years: A meta-analysis. Ecotox. Environ. Safe. 208, 111644.
Li, J., Huang, J., Cao, R., Yin, P., Wang, L., Liu, Y., Pan, X., Li, G., Zhou, M., 2021b. The association between ozone and years of life lost from stroke, 2013–2017: A retrospective regression analysis in 48 major Chinese cities. J. Hazard. Mater. 405, 124220.
Li, K., Jacob, D.J., Shen, L., Lu, X., De Smedt, I., Liao, H., 2020a. Increases in surface ozone pollution in China from 2013 to 2019: anthropogenic and meteorological influences. Atmos. Chem. Phys. 20, 11423–11433.
Li, R., Li Cui, L., Fu, H., Li, J.L., Zhao, Y., Chen, J., 2020b. Satellite-based estimation of full-coverage ozone (O3) concentration and health effect assessment across Hainan Island. J. Clean Prod. 244, 118773.
Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., Zhang, Q., He, K., 2017a. Anthropogenic emission inventories in China: a review. Natl. Sci. Rev. 4, 834–866.
Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017b. Estimating ground-level PM2.5 by fusing satellite and station observations: A geo-intelligent deep learning approach. Geophys. Res. Lett. 44, 11985–11993.
Li, T., Shen, H., Zeng, C., Yuan, Q., Zhang, L., 2017c. Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: Methods and assessment. Atmos. Environ. 152, 477–489.
Li, R., Zhao, Y., Zhou, W., Meng, Y., Zhang, Z., Fu, H., 2020c. Developing a novel hybrid model for the estimation of surface ozone (O3) across the remote Tibetan Plateau during 2005–2018. Atmos. Chem. Phys. 20, 6159–6175.
Liang, S., Li, X., Teng, Y., Fu, H., Chen, L., Mao, J., Zhang, H., Gao, S., Sun, Y., Ma, Z., Azzi, M., 2019. Estimation of health and economic benefits based on ozone exposure level with high spatial-temporal resolution by fusing satellite and station observations. Environ. Pollut. 255, 113267.
Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., Bi, J., 2020. Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. Environ. Int. 142, 105823.
Lu, X., Hong, J., Zhang, L., Cooper, O.R., Schultz, M.G., Xu, X., Wang, T., Gao, M., Zhao, Y., Zhang, Y., 2018. Severe surface ozone pollution in China: a global perspective. Environ. Sci. Technol. Lett. 5, 487–494.
Ma, Z., Zhang, X., Xu, J., Zhao, X., Meng, W., 2011. Characteristics of ozone vertical profile observed in the boundary layer around Beijing in autumn. J. Environ. Sci. 23, 1316–1324.
Maji, K.J., Ye, W.F., Arora, M., Nagendra, S.M.S., 2019. Ozone pollution in Chinese cities: Assessment of seasonal variation, health effects and economic burden. Environo. Pollut. 247, 792–801.
Manning, W.J. and v. Tiedemann, A., 1995. Climate change: Potential effects of increased atmospheric Carbon dioxide (CO2), ozone (O3), and ultraviolet-B (UV-B) radiation on plant diseases. Environo. Pollut. 88, 219-245.
Neidell, M., Kinney, P.L., 2010. Estimates of the association between ozone and asthma hospitalizations that account for behavioral responses to air quality information. Environ. Sci. Policy. 13, 97–103.
Peng, X., Shen, H., Zhang, L., Zeng, C., Yang, G., He, Z., 2016. Spatially continuous mapping of daily global ozone distribution (2004–2014) with the Aura OMI sensor. J. Geophys. Res.-Atmos. 121, 12702–12722.
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.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575.
Schauberger, B., Rolinski, S., Schaphoff, S., Müller, C., 2019. Global historical soybean and wheat yield loss estimates from ozone pollution considering water and temperature as modifying effects. Agr. Forest Meteorol. 265, 1–15.
Shen, H., Li, T., Yuan, Q., Zhang, L., 2018. Estimating regional ground-level PM2.5 directly from satellite top-of-atmosphere reflectance using deep belief networks. J. Geophys. Res.-Atmos. 123, 13875–13886.
Shen, H., Jiang, Y., Li, T., Cheng, Q., Zeng, C., Zhang, L., 2020. Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data. Remote Sens. Environ. 240, 111692.
Sicard, P., Crippa, P., De Marco, A., Castruccio, S., Giani, P., Cuesta, J., Paoletti, E., Feng, Z., Anav, A., 2021. High spatial resolution WRF-Chem model over Asia: Physics and chemistry evaluation. Atmos. Environ. 244, 118004.
Tian, Y., Wu, Y., Liu, H., Si, Y., Wu, Y., Wang, X., Wang, M., Wu, J., Chen, L., Wei, C., Wu, T., Gao, P., Hu, Y., 2020. The impact of ambient ozone pollution on pneumonia: A nationwide time-series analysis. Environ. Int. 136, 105498.
Travis, K.R., Jacob, D.J., 2019. Systematic bias in evaluating chemical transport models with maximum daily 8h average (MDA8) surface ozone for air quality applications: A case study with GEOS-Chem v9.02. Geosci. Model Dev. 12, 3641–3648.
Van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., Lyapustin, A., Sayer, A.M., Winker, D.M., 2016. Global estimates of fine particulate matter using a Combined Geophysical-Statistical method with information from satellites, models, and monitors. Environ. Sci. Technol. 50, 3762–3772.
Wang, H., Chai, S., Tang, X., Zhou, B., Bian, J., Vömel, H., Yu, K., Wang, W., 2019a. Verification of satellite ozone/temperature profile products and ozone effective height/temperature over Kunming. China. Sci. Total Environ. 661, 35–47.
Wang, C., Huang, X., Han, Y., Zhu, B., He, L., 2017. Sources and potential photochemical roles of formaldehyde in an urban atmosphere in South China. J. Geophys. Res.-Atmos. 122, 11934–11947.
Wang, W., Liu, X., Bi, J., Liu, Y., 2022. A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology. Environ. Int. 158, 106917.
Wang, Q., Miao, H., Warren, J.L., Ren, M., Benmarhnia, T., Knibbs, L.D., Zhang, H., Zhao, Q., Huang, C., 2021a. Association of maternal ozone exposure with term low birth weight and susceptible window identification. Environ. Int. 146, 106208.
Wang, J., Wang, S., Li, S., 2019b. Examining the spatially varying effects of factors on PM2.5 concentrations in Chinese cities using geographically weighted regression modeling. Environ. Pollut. 248, 792–803.
Wang, X., Wu, Z., Liang, G., 2009. WRF/CHEM modeling of impacts of weather conditions modified by urban expansion on secondary organic aerosol formation over Pearl River Delta. Particuology. 7, 384–391.
Wang, Y., Yuan, Q., Li, T., Zhu, L., Zhang, L., 2021b. Estimating daily full-coverage near surface O3, CO, and NO2 concentrations at a high spatial resolution over China based on S5P-TROPOMI and GEOS-FP. ISPRS J. Photogramm. Remote Sens. 175, 311–325.
Wei, J., Li, Z., Pinker, R.T., Wang, J., Sun, L., Xue, W., Li, R., Cribb, M., 2021. Himawari-8-derived diurnal variations in ground-level PM2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). Atmos. Chem. Phys. 21, 7863–7880.
Xie, Y.H., Fan, S.Y., Chen, M., Shi, J.C., Zhong, J.Q., Zhang, X.Y., 2019. An assessment of satellite radiance data assimilation in RMAPS. Remote Sens. 11 (1), 54.
Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357.
Xue, T., Zheng, Y., Geng, G., Xiao, Q., Meng, X., Wang, M., Li, X., Wu, N., Zhang, Q., Zhu, T., 2020. Estimating spatiotemporal variation in ambient ozone exposure during 2013–2017 using a Data-Fusion Model. Environ. Sci. Technol. 54, 14877–14888.
Yadav, R., Sahu, L.K., Beig, G., Jaaffrey, S.N.A., 2016. Role of long-range transport and local meteorology in seasonal variation of surface ozone and its precursors at an urban site in India. Atmos. Res. 176, 96–107.
Yang, Y., Liang, Z., Ruan, Z., Zhang, S., Zhao, Q., Lin, H., 2020. Estimating the attributable burden of preterm birth and low birth weight due to maternal ozone exposure in nine Chinese cities. Atmos. Environ. 222, 117169.
Zaveri, R.A., Peters, L.K., 1999. A new lumped structure photochemical mechanism for large-scale applications. J. Geophys. Res.-Atmos. 104, 30387–30415.
Zaveri, R.A., Easter, R.C., Fast, J.D., Peters, L.K., 2008. Model for simulating aerosol interactions and chemistry (mosaic). J. Geophys. Res.-Atmos. 113, D13204.
Zhan, Y., Luo, Y., Deng, X., Grieneisen, M.L., Zhang, M., Di, B., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut. 233, 464–473.
Zhang, L., Lee, C.S., Zhang, R., Chen, L., 2017. Spatial and temporal evaluation of long term trend (2005–2014) of OMI retrieved NO2 and SO2 concentrations in Henan Province. China. Atmos. Environ. 154, 151–166.
Zhang, Y., Wen, X., Jang, C., 2010. Simulating chemistry–aerosol–cloud–radiation–climate feedbacks over the continental U.S. using the online-coupled Weather Research Forecasting Model with chemistry (WRF/Chem). Atmos. Environ. 44, 3568–3582.
Zhang, Y., Wang, Y., Crawford, J., Cheng, Y., Li, J., 2018. Improve observation-based ground-level ozone spatial distribution by compositing satellite and surface observations: A simulation experiment. Atmos. Environ. 180, 226–233.
Zhang, X., Zhao, L., Cheng, M., Chen, D., 2020. Estimating ground-level ozone concentrations in Eastern China using satellite-based precursors. IEEE T. Geosci. Remote. 58, 4754–4763.
Zhao, T., Tesch, F., Markevych, I., Baumbach, C., Janßen, C., Schmitt, J., Romanos, M., Nowak, D., Heinrich, J., 2020. Depression and anxiety with exposure to ozone and particulate matter: An epidemiological claims data analysis. Int. J. Hyg. Envir. Heal. 228, 113562.
Zheng, B., Tong, D., Li, M., Liu, F., Hong, C., Geng, G., Li, H., Li, X., Peng, L., Qi, J., Yan, L., Zhang, Y., Zhao, H., Zheng, Y., He, K., Zhang, Q., 2018. Trends in China's anthropogenic emissions since 2010 as the consequence of clean air actions. Atmos. Chem. Phys. 18, 14095–14111.
Zhong, J.-Q., Lu, B., Wang, W., Huang, C.C., Yang, Y., 2020. Impact of soil moisture on winter 2m temperature forecasts in Northern China. J. Hydrometeorology 21, 597–614.