https://doi.org/10.5194/essd-2022-187
Preprint. Discussion started: 7 June 2022

Open Access Earth System Science Data Discussions

Reconstructed daily ground-level O_3 in China over 2005-2021 for climatological, ecological, and health research

Chenhong Zhou[1,*], Fan Wang[2,*], Yike Guo[1], Cheng Liu[3], Dongsheng Ji[4], Yuesi Wang[4], Xiaobin Xu[5], Xiao Lu[6], Yan Wang[7], Gregory R. Carmichael[8], Meng Gao[2]

[1]Department of Computer Science, Faculty of Science, Hong Kong Baptist University, Hong Kong SAR, China
[2]Department of Geography, Faculty of Social Sciences, Hong Kong Baptist University, Hong Kong SAR, China
[3]Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, 230026, Anhui Province, China
[4]State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, 100029, China
[5]State Key Laboratory of Severe Weather & Key Laboratory for Atmospheric Chemistry of China Meteorological Administration, Chinese Academy of Meteorological Sciences, Beijing, 100081, China
[6]School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai, 519082, Guangdong Province, China
[7]Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong SAR, China
[8]Department of Chemical and Biochemical Engineering, University of Iowa, Iowa City, IA 52242, USA
[*]These authors contributed equally to this work.

Correspondence to: Meng Gao (mmgao2@hkbu.edu.hk)

Abstract. Accompanied by the continuous declines of $PM_{2.5}$, O_3 pollution has become increasingly prominent and has been targeted by the Government of China to protect climate, ecosystem, and human health. Although satellite retrievals of column O_3 have been operated for decades and nationwide monitoring of ground-level O_3 has been offered since 2013 in China, climatological variability of ground-level O_3 remains unknown, which impedes understanding of the long-term driver and impacts of O_3 pollution in China. Here we develop an eXtreme Gradient Boosting (XGBoost) model integrating high-resolution meteorological data, satellite retrievals of trace gases, etc. to provide reconstructed daily ground-level O_3 over 2005-2021 in China. Model validation confirms the robustness of this dataset, with R^2 of 0.89 for sample-based cross-validation. The accuracy of the long-term variations has also been confirmed with independent historical observations covering the same period from urban, rural and background sites. Our dataset covers the long time period of 2005-2021 with 0.1°×0.1° gap-free grids, which can facilitate climatological, ecological, and health research. The dataset is freely available at Zenodo (https://zenodo.org/record/6507706#.Yo8hKujP13g; Zhou, 2022).

1 Introduction

Tropospheric ozone (O_3) is an important air pollutant (Fishman and Crutzen, 1978) and the third most potent greenhouse gas (IPCC, 2021). It also produces hydroxyl, a key oxidant in atmospheric chemistry, to modulate levels of secondary pollutants (Crutzen 1988; Gao et al., 2020). Many studies confirm that exposure to high surface O_3 levels is harmful to plant growth, crop yields, and human health (Sandermann, 1996; Tingey and Taylor, 1980; Heagle, 1989; Heck et al., 1982; Lippmann, 1989). With accelerated industrialization and urbanization since 2000 in China, emissions of O_3 precursors (nitrogen oxides and volatile organic compounds) have increased substantially (Kurokawa et al., 2013; Ohara et al., 2007; Zheng et al., 2018), leading to emerging and widespread O_3 pollution with threats to health and food security (Lu et al., 2018, 2020).

To better manage air quality, the Chinese government has operated a nationwide monitoring network since 2013 (Liu et al., 2021) to measure concentrations of six air pollutants, including O_3. In the same year, the Chinese government implemented a series of strengthened emission control measures to tackle the rising public discontent with poor air quality. Since then, $PM_{2.5}$ has exhibited a persistent decrease, yet an increasing trend of O_3 has been observed (Li et al., 2019; Lu et al., 2018). This nationwide network makes it possible to understand recent trends and causes of O_3 pollution in urban regions of China. However, most of the sites of the network have been distributed in major populated centers and in urban areas. The unequal allocation of monitoring sites hinders comprehensive assessments of the impacts of O_3 on human health and the ecosystem, particularly in rural and less-settled areas (Liu et al., 2020). Besides, the long-term variations of ground-level O_3 in many regions remain unknown, leading to difficulties in understanding how climate variability influences ground-level O_3 in China.

To fill these gaps, satellite retrievals of trace gases have been used to supplement surface observations as satellite monitoring has a longer history and global coverage (Ziemke et al., 2019). Unfortunately, it is challenging to infer near-surface O_3 using satellite column O_3 retrievals, as near-surface O_3 typically occupies only a few percent and there is a weak correlation between ground-level O_3 and satellite-retrieved O_3 (Hayashida et al., 2018; Liu et al., 2020; Wei et al., 2022; Zhang et al., 2020; Shen et al., 2019). Consequently, accurate and long-term near-surface O_3 concentrations cannot be derived from satellite measurements alone. To estimate surface O_3 concentrations with complete spatiotemporal coverage, statistical methods have been widely used (Chen et al., 2021; Kerckhoffs et al., 2015; Qiao et al., 2019; Zhang et al., 2020), including both traditional statistical methods and machine learning algorithms. Traditional statistical models, such as land-use regression (LUR) and geographically weighted regression (GWR), were built to explore relationships between surface O_3 and potentially influencing factors, including O_3 precursors, land-use types, elevation, and meteorological variables (Liu et al., 2020; Wei et al., 2022). For example, Chen et al. (2021) developed a hybrid land-use regression (LUR) and Bayesian Maximum Entropy (BME) to predict daily surface O_3 concentrations in China over 2015-2017. Zhang et al. (2020) proposed a geographically weighted regression (GWR) model with spatial information embed into the linear regression model to predict O_3 concentrations. Generally, most of these traditional statistical methods are based on linear regression, which is not ideal to handle the non-linearity and high-order interactions between predictors and O_3.

https://doi.org/10.5194/essd-2022-187
Preprint. Discussion started: 7 June 2022

Machine learning algorithms are capable of mining and fitting nonlinear relationships from big data. Multiple machine learning algorithms, including neural network (Di et al., 2017), random forests (RF) (Li et al., 2020; Zhan et al., 2018), extreme gradient boosting (XGBoost) (Liu et al., 2020), and extremely randomized trees (ERT) (Wei et al., 2022), have been adopted to predict ground-level O_3, and superior performances have been achieved compared to traditional statistical models. Most of these related studies presented results over a relatively short time period due to lack of ground-level observations. Liu et al. (2020) used XGBoost model and independent validation to offer nationwide daily estimations of O_3 in China from 2005 to 2017. However, R^2 values of cross-validation were generally lower than 0.80. Considering the spatiotemporal heterogeneity of O_3, Wei et al. (2022) employed an extended ERT model, named space-time extremely randomized trees (STET), to integrate space and time information with other independent variables during model training, which enhanced the performance of prediction.

With a thorough review of existing studies of surface O_3 estimation, we found most of them more or less suffer from drawbacks in terms of coarse spatial resolution, short-covering period, or low prediction accuracy. To facilitate climatological, ecological, and health research of O_3, we aim to provide a long-term, full-coverage, and high-resolution daily ground-level O_3 across China. We trained a nationwide ozone prediction model based on the XGBoost algorithm using atmospheric reanalysis, remote sensing products, emission estimates, etc. as predictors, and ground-based observations from 2013 to 2021 as predictands. Both sample-based and station-based cross-validation methods were conducted to evaluate the spatial and temporal predictive ability of the model. Then we used the well-trained model to generate long-term and full-coverage ground-level O_3 concentrations in China for a 17-year-long period from 2005 to 2021 at a spatial resolution of 0.1° × 0.1°. To test the model generalization performance, we conducted independent validation over 2005-2021 by comparing the generated O_3 predictions with observations from several stations covering urban, rural and background areas.

2 Materials and methods

2.1 Meteorological and air pollution data

Table 1 summarizes the ground-based observations, atmospheric reanalysis, satellite remote sensing products, and anthropogenic emission estimates used in this study. Hourly observations of ground-level O_3 across mainland China from the year 2013 to 2021 were obtained from the China National Environmental Monitoring Center (CNEMC) network (http://www.cnemc.cn/en/). It started from ~900 monitoring stations in 2013 to ~1600 in 2021. We removed negative O_3 values and then calculated the daily maximum 8 h average (MDA8) ozone concentrations for each monitoring site. As the abundance of O_3 in the troposphere is affected by both emissions (anthropogenic and natural) and meteorological conditions (Gao et al., 2016; Jacob and Winner, 2009), we considered meteorological variables, anthropogenic emission inventory, elevation, land use, normalized difference vegetation index (NDVI), etc. as input variables for the machine learning model (Table 1).

Nine considered meteorological variables include downward shortwave radiation (DSR), near-surface air temperature (TEM), relative humidity (RH), surface pressure (SP), boundary layer height (BLH), precipitation (PRE), evaporation (ET) and winds (horizontal and vertical components, WU and WV). DSR, TEM, RH, and ET affect the photochemical production of O_3, while SP, BLH, PRE, and winds are essential for the transport, diffusion, and deposition of O_3 (Jacob and Winner 2009). We obtained these meteorological variables from the hourly ERA5 reanalysis dataset (Hersbach et al., 2020) at a high spatial resolution of 0.1° × 0.1°. Daily column concentrations of tropospheric O_3 and NO_2 were recorded by the Ozone Monitoring Instrument (OMI) onboard NASA's Aura satellite and offered at a relatively coarser spatial resolution of 0.25° × 0.25°. Land use cover (LUC) and NDVI products were collected from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite retrievals (Justice et al., 2002), while digital elevation model (DEM) data were taken from the Shuttle Radar Topography Mission (SRTM) dataset (Reuter, Nelson, and Jarvis 2007).

Emissions of major O_3 precursors, namely nitrogen oxides (NOx), volatile organic compounds (VOCs), and carbon monoxide (CO), over 2005-2020 were provided by the Multiresolution Emission Inventory for China (MEIC) (Zheng et al., 2018; Zheng, Cheng, et al., 2021; Zheng, Zhang, et al., 2021). As the emission estimates for 2021 are not currently available from MEIC, we took the average of estimates for 2019 and 2020 as the conditions for 2021. Similarly, we used LUC in 2020 for 2021 considering the negligible change of LUC over such a short period of time. To deal with the mismatch of spatial resolution of these inputs, we followed previous studies (Liu et al., 2020; Wei et al., 2022; Zhan et al., 2018; Zhang et al., 2020) and used the bilinear interpolation to resample satellite O_3 and NO_2, and MEIC emission estimates at a coarser resolution to 0.1° × 0.1° grids. All these variables were converted into daily data to build the model at daily resolution.

2.2 Construction of a machine learning model

2.2.1 Spatiotemporal terms

Previous studies have shown that spatiotemporal heterogeneities are valuable to characterize the variations in O_3 but are usually neglected in most regression and machine learning explorations (Liu et al., 2020; Wei et al., 2020; Wei et al., 2022; Wei et al., 2021). To capture the relationship between O_3 concentrations and covariates that change over time and space, we introduced spatial and temporal information as additional input variables for model training. Similar to previous studies (Liu et al., 2020; Wei et al., 2020; Wei et al., 2022; Wei et al., 2021), we represented spatial information of one point in grid space as the Haversine great-circle distances between its longitude and latitude to the four corners and the center of the study region (i.e., 70.0°E-135.0°E, 15.0°N- 55.0°N). The time term was expressed by the day of the year (DOY) to capture O_3 variations along with seasonal cycles. These two spatiotemporal terms can account for spatial non-stationarity and spatiotemporal autocorrelations of O_3.

https://doi.org/10.5194/essd-2022-187
Preprint. Discussion started: 7 June 2022

2.2.2 Feature selection

Different variables play more or less influences on prediction, and we evaluated the importance of each independent variable for ground-level O_3 estimation. XGBoost library in Python was used to automatically measure the importance score for each variable and provide estimates of feature importance for the model decision-making process. The importance score indicates how useful or valuable each feature/variable is during model construction. A higher importance score suggests a more predominant role of the feature in making key decisions with decision trees. We show the feature importance of predictors in Figure 1, which also displays the changes of feature importance brought by introduction of the spatiotemporal terms. All 17 independent variables have importance scores >1% in Figure 1a, and we thus kept these variables in the model. In Figure 1b, we summarized the importance scores of five Haversine distances in the space term to explicitly indicate the contribution of spatial information. We found that DSR, Space, TEM, ET, and Time are the five most important variables for predicting daily O_3, with importance scores of 24.8%, 22.1%, 10.0%, 5.9%, and 4.4%, respectively. This also demonstrates that introducing spatiotemporal terms is valuable and necessary. Other meteorological variables and three main O_3 precursors influence the model to some degree, with importance scores ranging from 2% to 5%. OMI NO_2 and O_3 column products exhibit the least impact on O_3 estimates.

2.2.3 XGBoost modeling

We employed the extreme gradient boosting (XGBoost) (Chen and Guestrin 2016) algorithm to predict ground-level ozone (O_3) concentrations using a set of related predictor variables. XGBoost is a highly efficient machine learning algorithm based on gradient tree boosting and has been widely applied in many tasks. Previously, we adopted it to correct systematic bias of chemical transport model (Yin et al., 2021). XGBoost is one of the ensemble learning techniques that combine several weak models (e.g., decision trees) to generate a strong model for better performance. The combination ways in ensemble learning have three main classes: bagging, stacking, and boosting. Different from random forests (RF) that generate predictions by averaging predictions from all the individual trees (tree bagging models), XGBoost, as a scalable tree boosting model, can capture the nonlinearity feature by constructing the weighted ensemble of weak prediction models. Due to its regularized boosting and parallel processing nature, XGBoost algorithm is not prone to overfitting and shows superiority in both speed and performance compared with other tree bagging models.

In this study, the open-source Python package XGBoost was used and we tuned hyperparameters of the model by a grid search to obtain optimal hyperparameter values and considerably narrowed search space (Liu et al., 2020; Xiao et al., 2018). Specifically, three hyperparameters: namely the number of trees (n_estimators), the maximum depth of the tree (max_depth), and the sample weight of the smallest leaf node (min_child_weight), exhibited great effects on model performance in our experiments. We thus took them as tunable parameters and adjusted them carefully. After determining the optimal form of hyperparameters, we trained the XGBoost model using all available training data to generate a long-term, full-coverage, and high-resolution daily O_3 in China from 2005 to 2021, and flowchart is shown in Figure 2.

2.3 Model validation

To evaluate the overall model performance in estimating daily MDA8 O_3 concentrations, we adopted both sample-based (out-of-sample) and station-based (out-of-station) 10-fold cross-validation (CV). In the sample-based CV scheme, all data samples were randomly divided into 10 groups where 1 group was selected as the validation data and the rest 9 groups were used for model training. This operation was repeated 10 times until all samples had been tested (Molinaro, Simon, and Pfeiffer 2005). The data in a station-based CV scheme were separated by their locations, and training and validation samples did not overlap in space. Therefore, the station-based CV scheme was capable of testing the estimation accuracy at locations where ground-based O_3 measurements were not available during training.

We used multiple statistical indicators, including coefficient of determination (R2), root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percent error (MAPE), and ordinary least squares (OLS) regression (e.g., slope and intercept) to present the estimation accuracy of the model at the daily and monthly scales. We also used observations of MDA8 O_3 from independent stations (stations not used in training) and over a longer time period (beyond the training time period) to demonstrate the reliability of our predictions. Independent observations of MDA8 O_3 at three stations were taken from Xu et al. (2020), and these stations are being maintained by China Meteorological Administration (CMA) as a part of the Global Atmosphere Watch (GAW) programme. We used a rural site (Gucheng), an urban site (CMA campus, CMA) and a background site (Shangdianzi) in the comparison. The locations and more descriptions of these sites are documented in Xu et al. (2020).

3 Results and discussion

3.1 Sample-based cross-validation

Table 2 displays cross-validation results for each year and for years of 2013-2021. The total number of data samples over 2013-2021 across China is over 3.9 million (N=3,957,573). The R2, RMSE, MAE, and MAPE of sample-based CV are 0.89, 15.22 µg/m3, 10.18 µg/m3, and 17.82%, respectively. This indicates that our predicted daily MDA8 O_3 concentrations are in good agreement with ground-measured O_3 concentrations, with the slope and y-intercept from linear regression equal to 0.86 and 12.25 µg/m3, respectively (Figure 3a). The R^2 values of sample-based CV for each year exhibit continuous increases and the estimation errors (i.e., RMSE, MAE, MAPE) decrease over the period of 2013-2021. This is associated with the increased number of monitoring stations built in recent years in China and the better quality-control of data (Wei et al., 2021, 2022). Except for 2013, the R2 values for 2014-2021 are generally over 0.87 (Table 2). The slightly worse performance for 2013 is due to the significantly lower number of data samples.

We also validated model performance for five populated and economically dynamic regions, namely the Beijing-Tianjin-Hebei (BTH), the North China Plain (NCP), the Yangtze River Delta (YRD), the Pearl River Delta (PRD), and the Sichuan Basin (SCB) regions. According to the definitions of these five regions in (Lu et al., 2019), there are 90, 308, 230, 57, and 99 monitoring stations located in the BTH, NCP, YRD, PRD, and SCB, respectively. Figure 3b-f shows that all the R2 values are generally over 0.88 and all the slopes are close to 1.0. Besides, small prediction errors (e.g., RMSE = 14.8–17.1 µg/m3, MAE = 9.1–11.2 µg/m3, and MAPE = 14.9–19.6%) have been achieved.

Considering the spatial inhomogeneity of O_3 monitoring stations, we performed an additional validation on the individual-station scale, and performance for each monitoring station is displayed in Figure 4. We can find that stations with high estimation accuracy are mostly located in eastern China, especially in the BTH, NCP, and YRD regions, which agree with the results shown in Figure 3b-d. The stations with slightly worse performance are generally located in the northeastern and northwestern China, which is associated with the lower number of stations there. In general, ~ 85% of the stations have R2 values exceeding 0.8, RMSE < 15 µg/m3, MAE ~ 10 µg/m3, and MAPE < 18%. When we evaluated the predicted monthly mean MDA8 O_3 estimates from 2013 to 2021, we found notable improvements, with R2 = 0.92, RMSE = 9.45 µg/m3, and MAE = 5.34 µg/m3 (Figure S1). Similarly, relatively poor performances are seen for 2013, and higher R2 values of 0.86–0.97 are found for other years (Figure S1).

3.2 Station-based cross-validation and Independent validation

The station-based CV results are slightly worse than those of sample-based, whereas R2 values still reach 0.79 (Table 2, Figure S2). The spatial predictive performance is also observed to gradually increase over years (Table 2). O_3 pollution has distinct contrasts in different regions of China, leading to notable differences between training and validation samples in station-based CV. Degraded performance in station-based CV was also found in previous studies (Liu et al., 2020; Wei et al., 2022). Due to different densities of monitoring stations, predictive ability varies in the five concerned regions (Figure S2b-f). The performances are better in BTH and NCP (R2 ~ 0.85) than those in the YRD, PRD, and SCB regions (R2 ~ 0.8). Besides, the model displays notable spatial differences in the spatial predictive ability, as shown in Figure S3. Better predictive ability is found for most stations in eastern and central China, with high R2 values > 0.8 and small RMSE, MAE, and MAPE values of < 18 µg/m3, 12 µg/m3, and 18%, respectively. This is consistent with the results displayed in Figure S2b-f that more accurate predictions are found for the BTH and NCP regions.

To demonstrate the accuracy of this long-term dataset, we used O_3 observations from four stations that cover periods before 2013. As shown in Figure 5, reconstructed O_3 values are generally consistent with observations with respect to the temporal variations, with high R2 values ranging from 0.71 to 0.89 and RMSE varying between 17.38 and 22.13 µg/m3. The accuracy of this dataset is thus confirmed in urban, rural and background sites.

3.3 Comparison with related studies

A comprehensive comparison between this study and related studies was conducted and the results are listed in Table 3. CV-R2 and RMSE values from sample-based CV are reported. (Zhang et al., 2018) generated a nationwide O_3 dataset for China with the RF algorithm. However, data for only year 2015 was provided and a much lower R2 of 0.69 and a higher RMSE of 26.00 µg/m3 were achieved. With the GWR method, (Zhang et al., 2020) estimated monthly O_3 concentrations in 2014 based on satellite-based precursors, while their study achieved inferior performance (R2= 0.77) compared to our predictions on monthly scale (Figure S1b, R^2= 0.86 and RMSE = 14.17 μg/m3 for 2014). (Liu et al., 2020) slightly improved the prediction accuracy with R2 values of 0.78 based on the XGBoost algorithm and conducted independent validation of the data accuracy before 2013. Later on, Chen et al. (2021), Xue et al. (2020), and Wei et al. (2022) adopted other algorithms to improve the prediction accuracy or spatial resolution.

In this study, our model yielded a higher accuracy with R2 = 0.89 and RMSE = 15.22 μg/m3. The dataset covers unprecedent long term coverage, and the accuracy before 2013 was also validated although observations are rare. We compared the spatial distribution of predicted O_3 in this study and that from ChinaHighO3 provided by Wei et al. (2022). As other O_3 datasets are not accessible online, we do not show the comparison. As shown in Figure 6, ChinaHighO3 fails to capture O3 hotspots on June 18 of 2015 and August 13 of 2019, especially for the areas circled in red rectangular boxes. More inter-comparisons are provided in Figure S4.

3.4 Long-term spatiotemporal variations of surface O_3

By integrating a variety of predictors, including ground-based observations, satellite remote sensing data, emission inventory, and atmospheric reanalysis data, we employed the XGBoost model to generate a long-term full-coverage ground-level daily O_3 dataset in China, which covers the long time period of 2005–2021 with a fine spatial resolution of 0.1° × 0.1°. By taking advantage of the long-temporal coverage, full-coverage, and high-resolution of this dataset, we analyzed the seasonal variation and trends of surface O_3 pollution in China. As displayed in Figure 7, significant variations in surface O3 concentrations for different seasons are seen in northern, central and southern China. In particular, summertime mean MDA8 O_3 concentrations in NCP and BTH regions are extremely high (NCP: mean O_3 = 133.4 ± 12.6 μg/m3; BTH: mean O_3 = 135.9 ± 14.7 μg/m3), whereas wintertime mean O_3 concentrations are much lower (NCP: mean O_3 = 60.3 ± 7.0 μg/m3; BTH: mean O_3 = 58.7 ± 7.8 μg/m3). In general, severe O_3 pollution occurs mainly in northern and central China during the summertime, O_3 concentrations in the PRD peak in the autumn (Figure 7), which is consistent with our previous investigation using chemical transport modeling (Gao et al., 2021). Besides, we noticed that O_3 concentrations in Yunnan Province peak in spring. Lower O_3 concentrations are found in most regions of China in winter due to inhibited photochemistry (Gao et al., 2016b).

Figure 8 presents the trends of annual mean MDA8 O_3 during the period of 2005-2013 and 2014-2021, and the trends for four seasons are shown in Figure 9. Over the period of 2005-2013, annual mean MDA8 O_3 declined significantly over eastern China, particularly in the NCP and YRD region. This trend reversed after 2013, and the increasing severity of O_3 pollution over China has been well documented previously (Li et al., 2019; Lu et al., 2018, 2020). The contrast trends are mainly due to China's clean air actions initiated in 2013 (Liu and Wang 2020). Since 2013, the dramatic decreases in $PM_{2.5}$ led to increases in O_3 concentrations by enhanced photolysis rates (Liu and Wang 2020). We also found that the declines of O_3 over the period of 2005-2013 mainly happened in spring while the recent enhancement mainly occurred in summer and autumn (Figure 9). Our ongoing study using this dataset aims to elucidate the reasons.

4 Summary

O_3 has emerged as a pollutant of growing concern in China, posing threats to both human health and ecosystem. Due to uneven and limited ground-based observations, the nationwide spatiotemporal patterns of O_3 pollution, especially their historical records remain unclear. In this study, we developed a nationwide daily MDA8 O_3 prediction model based on the XGBoost algorithm, together with meteorological variables, remote sensing products, and emission inventory as input predictors and ground-based observations as predictand. In addition to the core input variables (i.e., downwelling solar surface radiation, air temperature, evaporation), we considered spatiotemporal information, which were demonstrated to account for a large proportion of importance in the model decision-making process. The model performance were evaluated at varying spatiotemporal scales, and it indicated that the model has a strong predictive power by yielding high prediction accuracy and small estimation uncertainties, i.e., sample-based (station-based) CV-R2, RMSE, MAE, and MAPE values of 0.89 (0.79), 15.22 (21.16) µg/m3, 10.18 (14.06) µg/m3, and 17.82 (25.24) %, respectively.

A long-term, full-coverage, and high-quality daily O_3 concentration dataset with a spatial resolution of 0.1° × 0.1° covering the period of 2005–2021 in China was generated using the final model. Compared to other O_3 datasets, our dataset shows superiority in terms of data accuracy, spatial coverage and resolution, and data time span. To assess the data accuracy beyond the training data, independent validation was conducted using observations of O_3 over 2005-2021 from four sites. These four sites are typical urban, rural and background stations, and high degree of consistency has been found. Overall, our dataset exhibits a high prediction accuracy and can be used to characterize long-term surface O_3 variations over space and time, which is of great significance for policy-making for pollution control and relevant climatological, ecological, and health research.

Data Availability

The dataset described in this article is available online at: https://zenodo.org/record/6507706#.Yo8hKujP13g (Zhou et al., 2022).

Author Contributions

All the authors were involved in the generation of the introduced dataset. CZ, YG and MG wrote the manuscript with contributions from all the other authors

Declaration of Competing Interest

The author has declared that neither they nor their co-authors have any competing interests.

Financial Support.

This study was supported by grants from National Natural Science Foundation of China (No. 42005084) and Research Grants Council of the Hong Kong Special Administrative Region, China (project no. HKBU22201820 and HKBU12202021).

References

Chen, L., Liang, S., Li, X., Mao, J., Gao, S., Zhang, H., Sun, Y., Vedal, S., Bai, Z., and Ma, Z.: A hybrid approach to estimating long-term and short-term exposure levels of ozone at the national scale in China using land use regression and Bayesian maximum entropy, Sci. Tot. Env., 752, 141780, 2021.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 10.1145/2939672.2939785, 2016.

Crutzen, P. J.: Tropospheric ozone: An overview, Tropospheric ozone, 3-32, 1988.

Di, Q., Rowland, S., Koutrakis, P., and Schwartz, J.: A hybrid model for spatially and temporally resolved ozone exposures in the continental United States, J. Air & Waste Man. Ass., 67, 39-52, 2017.

Fishman, J. and Crutzen, P. J.: The origin of ozone in the troposphere, Nature, 274, 855-858, 10.1038/274855a0, 1978.

Gao, M., Carmichael, G. R., Saide, P. E., Lu, Z., Yu, M., Streets, D. G., and Wang, Z.: Response of winter fine particulate matter concentrations to emission and meteorology changes in North China, Atmos. Chem. Phys., 16, 11837-11851, 2016.

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

Hayashida, S., Kajino, M., Deushi, M., Sekiyama, T. T., and Liu, X.: Seasonality of the lower tropospheric ozone over China observed by the Ozone Monitoring Instrument, Atmospheric Environment, 184, 244-253, 2018.

Heagle, A. S.: Ozone and crop yield, Annual review of phytopathology, 27, 397-423, 1989.

Heck, W. W., Taylor, O., Adams, R., Bingham, G., Miller, J., Preston, E., and Weinstein, L.: Assessment of crop loss from ozone, Journal of the Air Pollution Control Association, 32, 353-361, 1982.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., and Schepers, D.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999-2049, 2020.

Jacob, D. J. and Winner, D. A.: Effect of climate change on air quality, Atmospheric environment, 43, 51-63, 2009.

Justice, C., Townshend, J., Vermote, E., Masuoka, E., Wolfe, R., Saleous, N., Roy, D., and Morisette, J.: An overview of MODIS Land data processing and product status, Remote sensing of Environment, 83, 3-15, 2002.

Kerckhoffs, J., Wang, M., Meliefste, K., Malmqvist, E., Fischer, P., Janssen, N. A., Beelen, R., and Hoek, G.: A national fine spatial scale land-use regression model for ozone, Environmental research, 140, 440-448, 2015.

Kurokawa, J., Ohara, T., Morikawa, T., Hanayama, S., Janssens-Maenhout, G., Fukui, T., Kawashima, K., and Akimoto, H.: Emissions of air pollutants and greenhouse gases over Asian regions during 2000–2008: Regional Emission inventory in ASia (REAS) version 2, Atmos. Chem. Phys. 13, 11019-11058, 2013.

Li, K., Jacob, D. J., Liao, H., Shen, L., Zhang, Q., and Bates, K. H.: Anthropogenic drivers of 2013–2017 trends in summer surface ozone in China, Proceedings of the National Academy of Sciences, 116, 422-427, 2019.

Li, R., Zhao, Y., Zhou, W., Meng, Y., Zhang, Z., and Fu, H.: Developing a novel hybrid model for the estimation of surface 8 h ozone (O 3) across the remote Tibetan Plateau during 2005–2018, Atmos. Chem. Phys., 20, 6159-6175, 2020.

Lippmann, M.: Health effects of ozone a critical review, Japca, 39, 672-695, 1989.

Liu, C., Gao, M., Hu, Q., Brasseur, G. P., and Carmichael, G. R.: Stereoscopic monitoring: a promising strategy to advance diagnostic and prediction of air pollution, Bulletin of the American Meteorological Society, 102, E730-E737, 2021.

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., and Bi, J.: Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach, Environment international, 142, 105823, 2020.

Liu, Y. and Wang, T.: Worsening urban ozone pollution in China from 2013 to 2017 – Part 2: The effects of emission changes and implications for multi-pollutant control, Atmos. Chem. Phys., 20, 6323-6337, 10.5194/acp-20-6323-2020, 2020.

Lu, X., Zhang, L., Wang, X., Gao, M., Li, K., Zhang, Y., Yue, X., and Zhang, Y.: Rapid increases in warm-season surface ozone and resulting health impact in China since 2013, Environmental Science & Technology Letters, 7, 240-247, 2020.

Lu, X., Hong, J., Zhang, L., Cooper, O. R., Schultz, M. G., Xu, X., Wang, T., Gao, M., Zhao, Y., and Zhang, Y.: Severe surface ozone pollution in China: a global perspective, Environmental Science & Technology Letters, 5, 487-494, 2018.

Lu, X., Zhang, L., Chen, Y., Zhou, M., Zheng, B., Li, K., Liu, Y., Lin, J., Fu, T.-M., and Zhang, Q.: Exploring 2016–2017 surface ozone pollution over China: source contributions and meteorological influences, Atmos. Chem. Phys., 19, 8339-8361, 2019.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M.: Prediction error estimation: a comparison of resampling methods, Bioinformatics, 21, 3301-3307, 2005.

Ohara, T., Akimoto, H., Kurokawa, J.-i., Horii, N., Yamaji, K., Yan, X., and Hayasaka, T.: An Asian emission inventory of anthropogenic emission sources for the period 1980–2020, Atmos. Chem. Phys., 7, 4419-4444, 2007.

Qiao, X., Guo, H., Wang, P., Tang, Y., Ying, Q., Zhao, X., Deng, W., and Zhang, H.: Fine Particulate Matter and Ozone Pollution in the 18 Cities of the Sichuan Basin in Southwestern China: Model Performance and Characteristics, Aerosol and Air Quality Research, 19, 2308-2319, 10.4209/aaqr.2019.05.0235, 2019.

Reuter, H. I., Nelson, A., and Jarvis, A.: An evaluation of void-filling interpolation methods for SRTM data, International Journal of Geographical Information Science, 21, 983-1008, 2007.

Sandermann Jr, H.: Ozone and plant health, Annual review of phytopathology, 34, 347-366, 1996.

Shen, L., Jacob, D. J., Liu, X., Huang, G., Li, K., Liao, H., and Wang, T.: An evaluation of the ability of the Ozone Monitoring Instrument (OMI) to observe boundary layer ozone pollution across China: application to 2005–2017 ozone trends, Atmos. Chem. Phys., 19, 6551-6560, 10.5194/acp-19-6551-2019, 2019.

Tingey, D. and Taylor Jr, G. E.: Variation in plant response to ozone: a conceptual model of physiological events, Environmental Protection Agency, Corvallis, OR (USA). Environmental Research …, 1980.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., and Cribb, M.: Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, Remote Sensing of Environment, 252, 112136, 2021.

Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W., Sun, L., Guo, J., Peng, Y., Li, J., and Lyapustin, A.: Improved 1 km resolution PM 2.5 estimates across China using enhanced space–time extremely randomized trees, Atmos. Chem. Phys., 20, 3273-3289, 2020.

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

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An ensemble machine-learning model to predict historical PM2. 5 concentrations in China from satellite data, Environmental science & technology, 52, 13260-13269, 2018.

Yin, H., Lu, X., Sun, Y., Li, K., Gao, M., Zheng, B., and Liu, C.: Unprecedented decline in summertime surface ozone over eastern China in 2020 comparably attributable to anthropogenic emission reductions and meteorology, Environmental Research Letters, 16, 124069, 2021.

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

Zhang, X., Zhao, L., Cheng, M., and Chen, D.: Estimating ground-level ozone concentrations in eastern China using satellite-based precursors, IEEE Transactions on Geoscience and Remote Sensing, 58, 4754-4763, 2020.

Zheng, B., Zhang, Q., Geng, G., Chen, C., Shi, Q., Cui, M., Lei, Y., and He, K.: Changes in China's anthropogenic emissions and air quality during the COVID-19 pandemic in 2020, Earth System Science Data, 13, 2895-2907, 2021a.

Zheng, B., Cheng, J., Geng, G., Wang, X., Li, M., Shi, Q., Qi, J., Lei, Y., Zhang, Q., and He, K.: Mapping anthropogenic emissions in China at 1 km spatial resolution and its application in air quality modeling, Science Bulletin, 66, 612-620, 2021b.

Zheng, B., Tong, D., Li, M., Liu, F., Hong, C., Geng, G., Li, H., Li, X., Peng, L., and Qi, J.: Trends in China's anthropogenic emissions since 2010 as the consequence of clean air actions, Atmos. Chem. Phys., 18, 14095-14111, 2018.

Ziemke, J. R., Oman, L. D., Strode, S. A., Douglass, A. R., Olsen, M. A., McPeters, R. D., Bhartia, P. K., Froidevaux, L., Labow, G. J., and Witte, J. C.: Trends in global tropospheric ozone inferred from a composite record of TOMS/OMI/MLS/OMPS satellite measurements and the MERRA-2 GMI simulation, Atmospheric Chemistry and Physics, 19, 3257-3269, 2019.

Zhou, C., Wang, F., Guo, Y., Carmichael, G.R., Liu, C., Wang, Y. and Gao, M.: Reconstructed daily ground-level MDA8 O3 over 2005-2021 in China, https://doi.org/10.5281/zenodo.6507706, 2022.

https://doi.org/10.5194/essd-2022-187
Preprint. Discussion started: 7 June 2022

380 **Table 1.** Summary of the data sources used in this study.

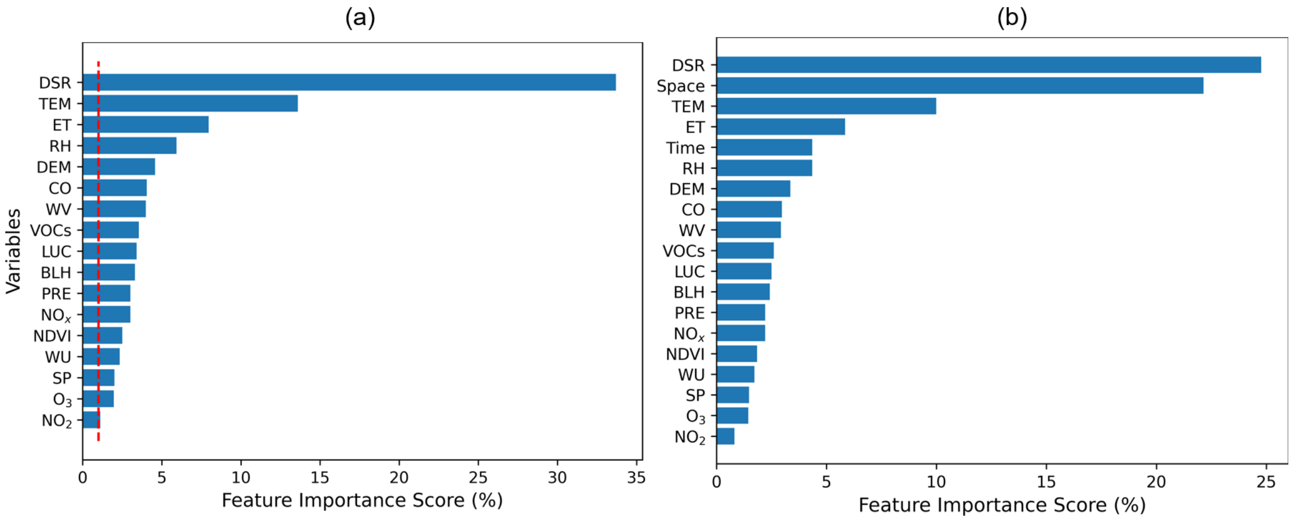
Category	Variables	Description	Units	Time range	Spatial resolution	Temporal resolution	Data source
Ground measurement	O_3	Ozone	μg/m^3	2013-2021	-	Hourly	CNEMC
Atmospheric reanalysis	DSR	Downwelling surface radiation	W/m^2	2005-2021	0.1°×0.1°	Hourly	ERA5 Land
	ET	Evaporation	mm				
	PRE	Precipitation	mm				
	RH	Relative humidity	%				
	TEM	2-m air temperature	K				
	SP	Surface pressure	hPa				
	WU	10-m u-component	m/s				
	WV	10-m v-component	m/s				
	BLH	Boundary layer height	m				
Satellite remote sensing products	O_3	tropospheric O_3	DU	2005-2021	0.25°×0.25°	Daily	OMI/Aura products
	NO_2	tropospheric NO_2	molecule/cm^2				
	NDVI	Normalized difference vegetation index	-	2005-2021	0.1°×0.1°	Monthly	MODIS products
	LUC	Land-use cover	-	2005-2020	0.1°×0.1°	Annual	
	DEM	Surface elevation	m	-	0.1°×0.1°	-	SRTM
Emission inventory	NO_x	Nitrogen oxides	Mg/grid	2005-2020	0.25°×0.25°	Monthly	MEIC
	VOCs	Volatile organic compounds					
	CO	Carbon monoxide					

Table 2. Cross-validation results of MDA8 O_3 estimation (μg/m^3) for each year from 2013 to 2021 in China.

Year	Sample size	Sample-based cross-validation				Station-based cross-validation			
	N	R^2	RMSE	MAE	MAPE	R^2	RMSE	MAE	MAPE
2013	193,613	0.79	23.34	14.37	38.08	0.53	35.27	23.83	65.77
2014	301,914	0.87	18.00	11.59	25.72	0.65	29.34	20.67	49.67
2015	504,888	0.88	16.46	10.48	21.03	0.66	27.07	19.06	40.07
2016	490,097	0.89	14.98	9.70	16.24	0.73	22.84	16.33	28.24
2017	512,664	0.93	12.41	8.28	11.62	0.86	17.80	12.60	17.13
2018	515,538	0.94	11.34	7.45	10.73	0.90	15.07	10.49	14.63
2019	402,519	0.95	10.69	7.03	9.51	0.91	14.26	9.89	13.02
2020	530,003	0.95	9.24	5.99	8.54	0.92	11.70	7.98	11.18
2021	506,337	0.95	9.07	5.94	8.38	0.92	11.17	7.67	10.67
All	3,957,573	0.89	15.22	10.18	17.82	0.79	21.16	14.06	25.24

Table 3. Comparison of the data quality of this study with other related studies focused on China.

Sources	Model	Training data period	Temporal resolution	Spatial resolution	CV-R^2	RMSE (μg m^{-3})	Independent validation	Products time range
Zhan et al. (2018)	RF	2015	Daily	0.1° × 0.1°	0.69	26.00	No	2015
Zhang et al. (2020)	GWR	2014	Monthly	0.25◦ × 0.25◦	0.77	-	No	2014
Liu et al. (2020)	XGBoost	2013–2017	Daily	0.1° × 0.1°	0.78	21.47	Yes	2005-2017
Xue et al. (2020)	Data fusion	2013–2017	Daily	0.1° × 0.1°	0.70	26.20	No	2013–2017
Chen et al. (2021)	LUR/BME	2015–2017	Daily	1km × 1km	0.80	23.50	No	2015–2017
Chen et al. (2016)	Iterative RF	2014–2019	Daily	0.0625◦	0.84	18.4	No	2008–2019
Wei et al. (2022)	STET	2013–2020	Daily	0.1° × 0.1°	0.87	17.10	No	2013–2020
This study	XGBoost	2013–2021	Daily	0.1° × 0.1°	0.89	15.22	Yes	2005-2021



400 **Figure 1.** Importance scores of variables before (a) and after adding spatiotemporal terms (b).

405

https://doi.org/10.5194/essd-2022-187
Preprint. Discussion started: 7 June 2022

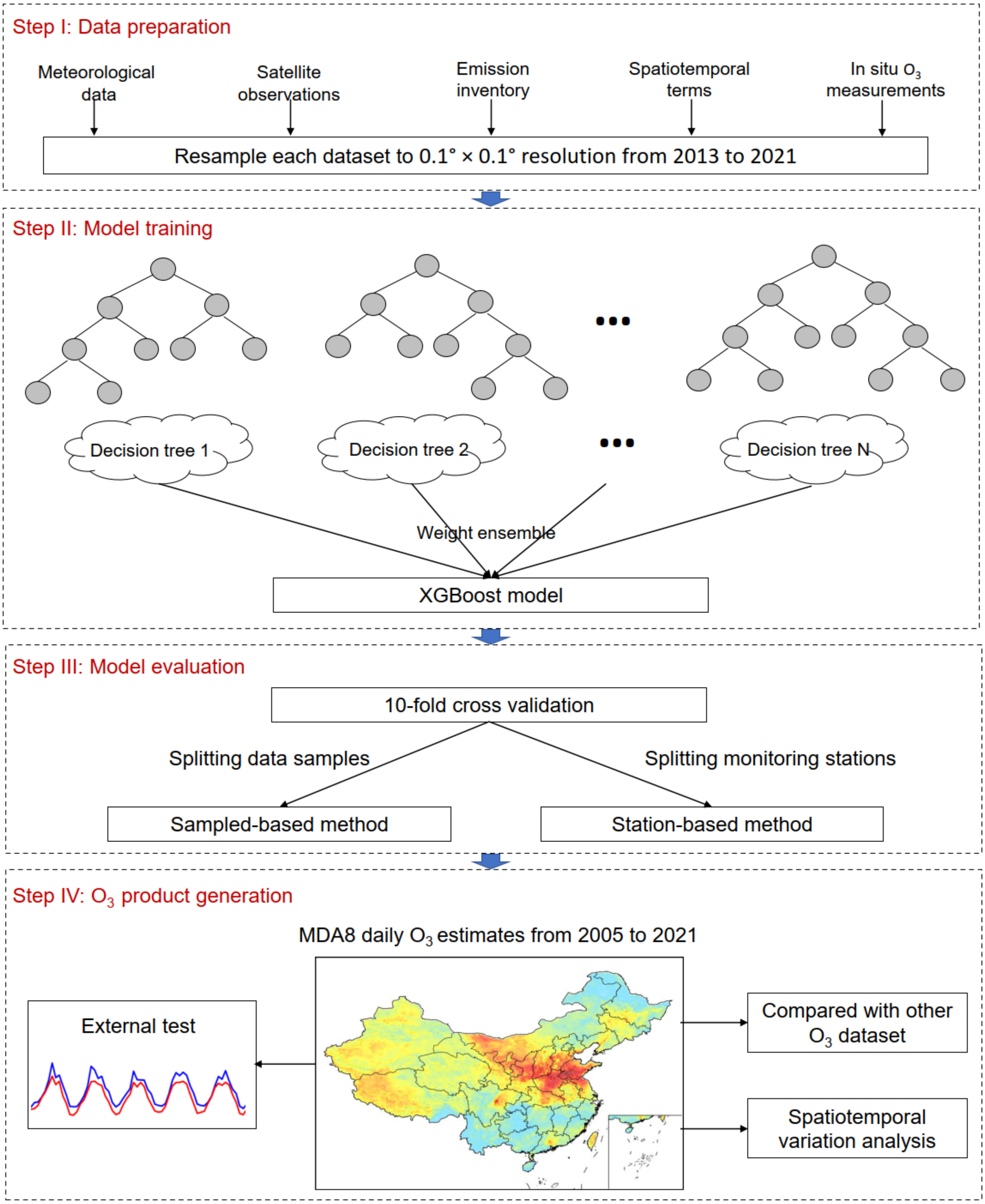


Figure 2. Flowchart for generating daily MDA8 O_3 dataset using the XGBoost model.

410

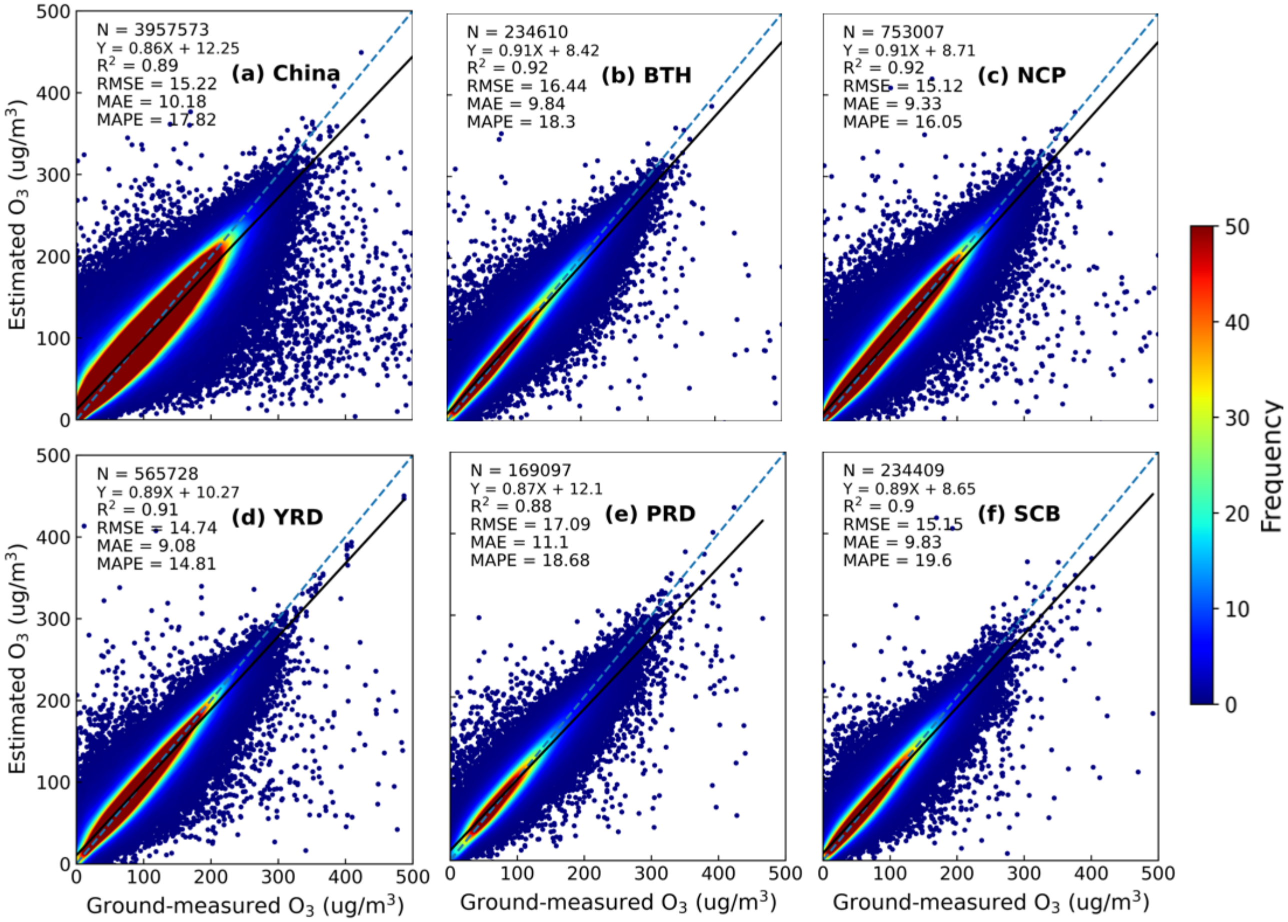
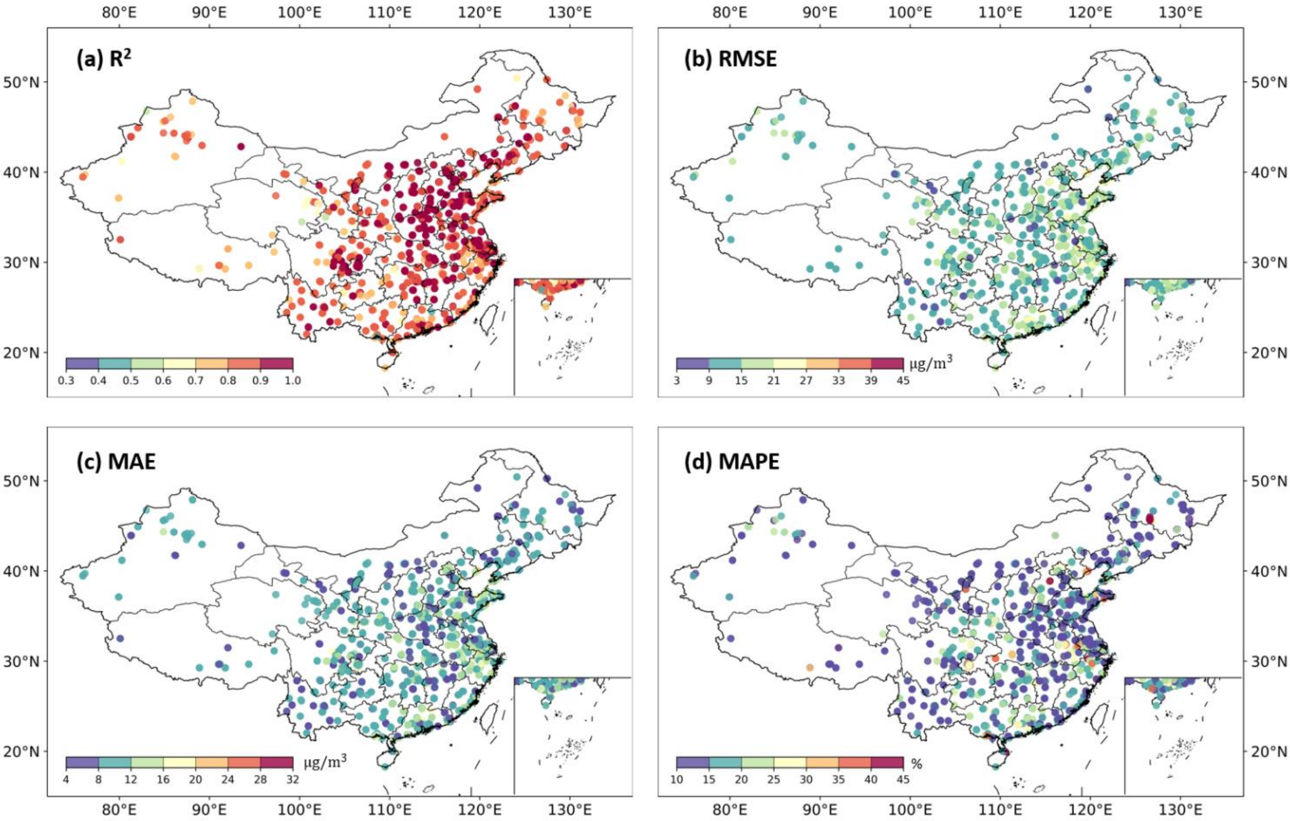


Figure 3. Density scatterplots of the sample-based cross-validation results of O_3 estimates from 2013 to 2021 (a) in China, (b) the Beijing-Tianjin-Hebei (BTH), (c) the North China Plain (NCP), (d) the Yangtze River Delta (YRD), (e) the Pearl River Delta (PRD), and (f) the Sichuan Basin (SCB).

415



420 **Figure 4.** Individual-station-scale sample-based cross-validation results of O_3 estimates from 2013 to 2021 in China.

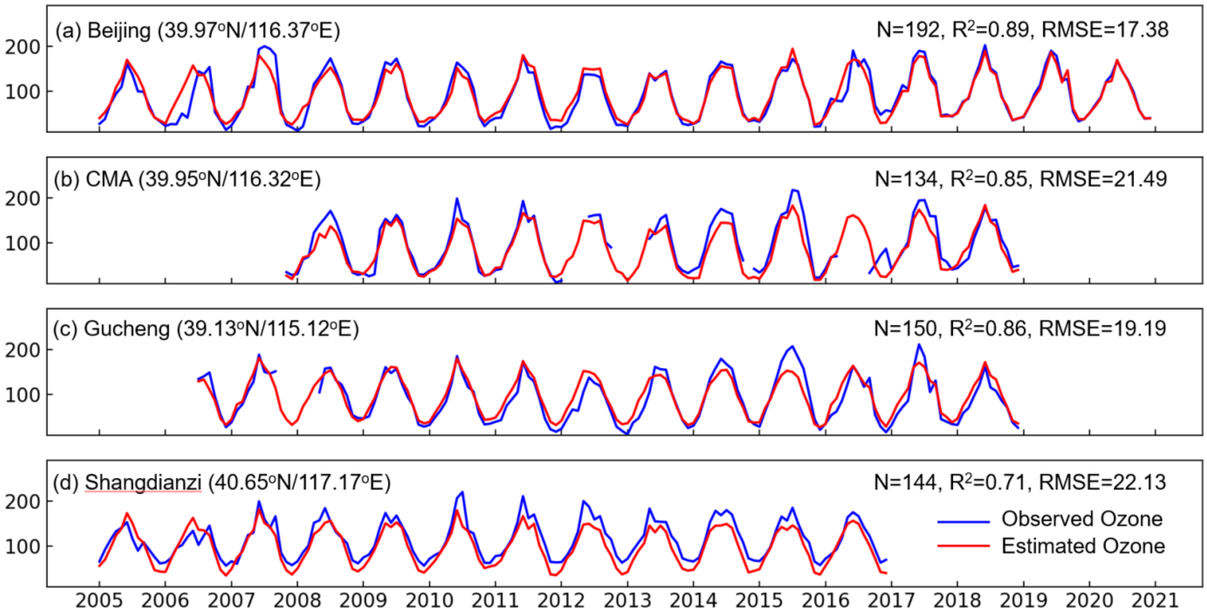


Figure 5. Time series of observed and estimated monthly mean MDA8 O_3 at four independent stations: Beijing IAP (Institute of Atmospheric Physics, Chinese Academy of Sciences) site, CMA (China Meteorological Administration) site, and Shangdianzi site, and Gucheng site in Hebei Province. Exact latitude and longitude information of each station is also provided.

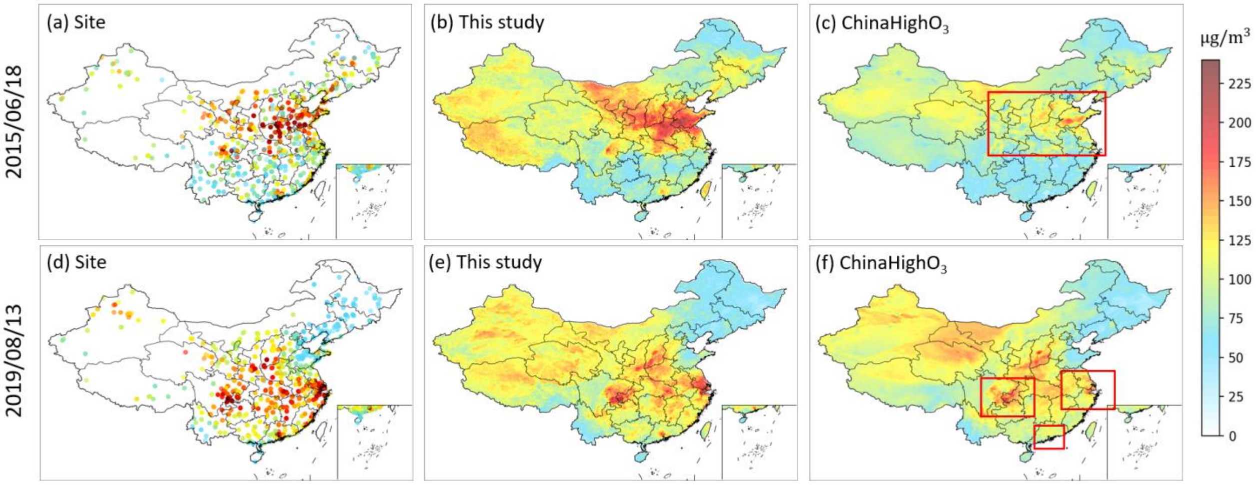


Figure 6. Comparison of O_3 distribution. From the left to right, it shows in situ O_3 concentration measurements, the O_3 dataset generated in this study, and ChinaHighO_3, respectively. Dates are given in the format year/month/day.

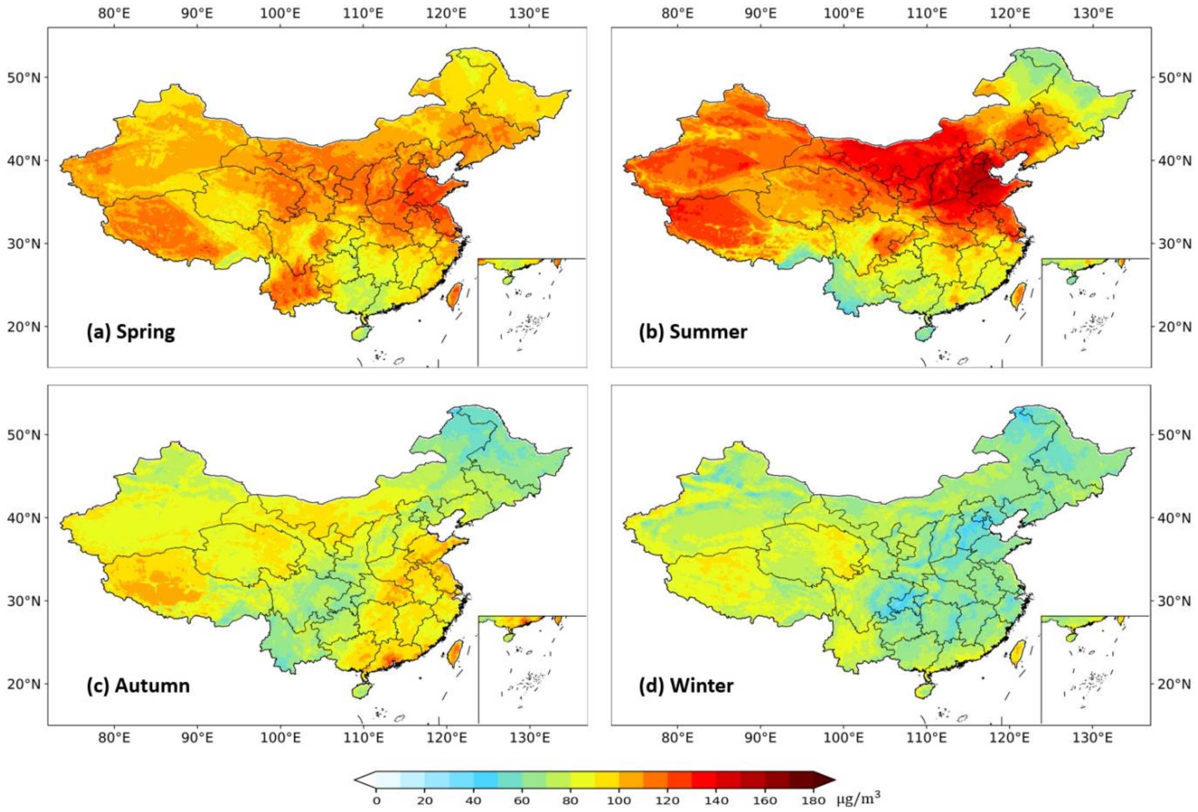

(a) Spring
(b) Summer
(c) Autumn
(d) Winter
80°E 90°E 100°E 110°E 120°E 130°E
50°N 40°N 30°N 20°N
0 20 40 60 80 100 120 140 160 180 μg/m³

Figure 7. Multi-year (2005–2021) seasonal mean MDA8 O_3 distributions across China.

440

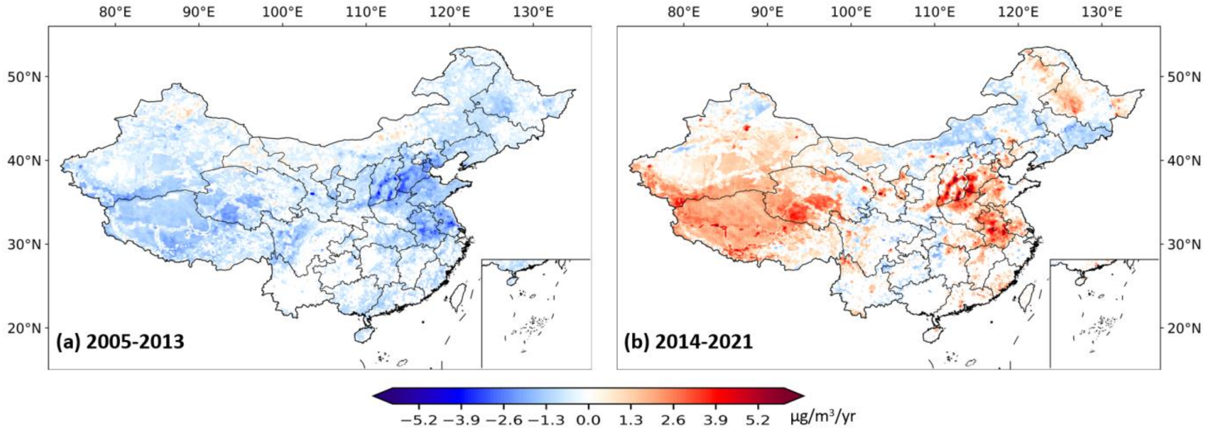
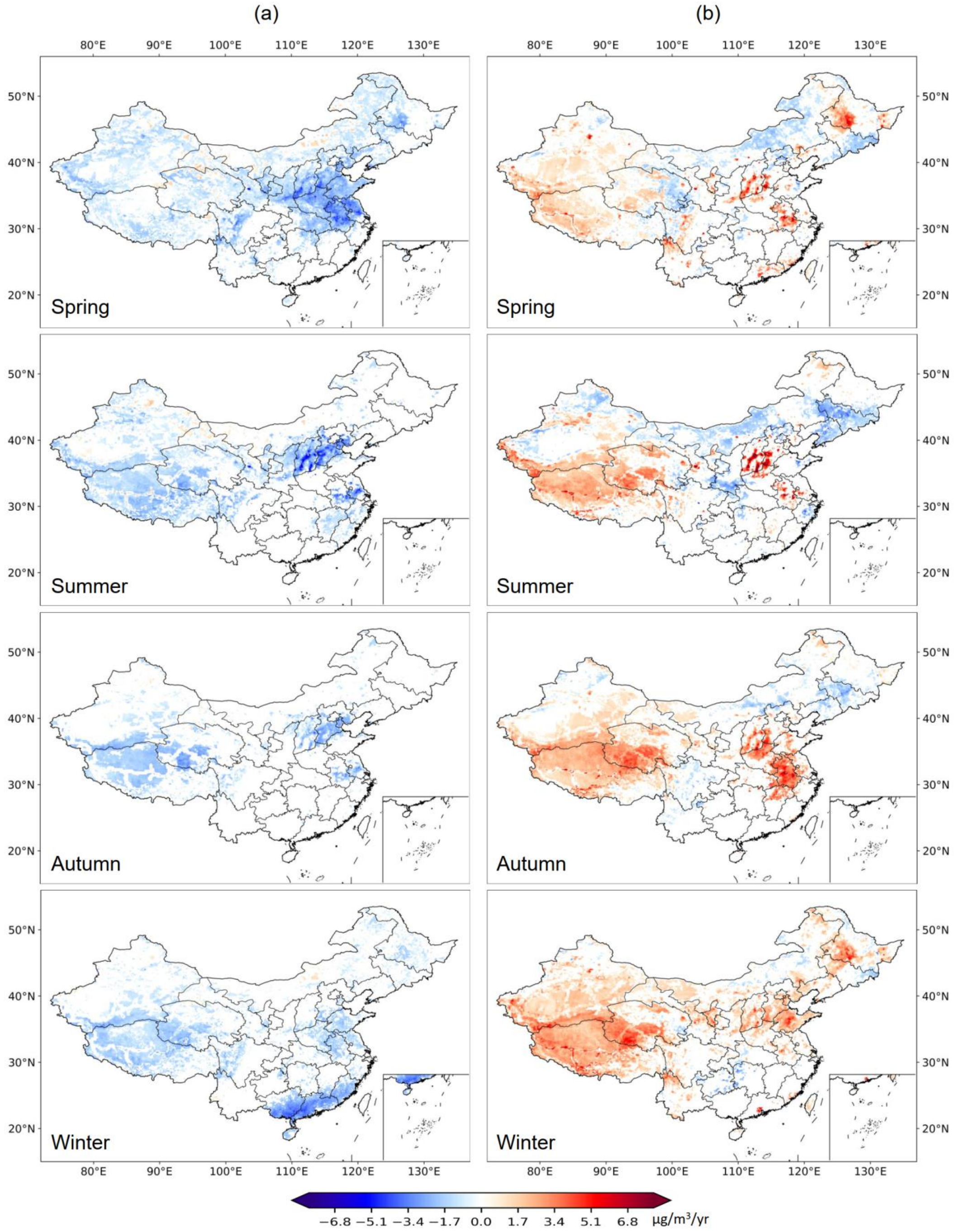

(a) 2005-2013
(b) 2014-2021
80°E 90°E 100°E 110°E 120°E 130°E
50°N 40°N 30°N 20°N
−5.2 −3.9 −2.6 −1.3 0.0 1.3 2.6 3.9 5.2 μg/m³/yr

Figure 8. O_3 variation trends during the period of 2005-2013 (a) and 2014-2021 (b). Insignificant trends (p>0.05) are not shown.

https://doi.org/10.5194/essd-2022-187
Preprint. Discussion started: 7 June 2022



445 **Figure 9.** Trends of seasonal mean MDA8 O_3 during the period of 2005-2013 (column (a)) and 2014-2021 (column (b)). Insignificant trends ($p>0.05$) are not shown.