Supporting information

Text S1. Data and Methodology for long-term ozone modeling

S1.1 Data used for ground-level ozone modeling

In-situ ozone monitoring data

Hourly ground-level ozone measurements from 2013 to 2020 were collected across mainland China, sourced from the China National Environmental Monitoring Center. Additional hourly measurements were acquired from the Environmental Protection Departments of the Hong Kong (https://cd.epic.epd.gov.hk/EPICDI) and Macau (https://www.dspa.gov.mo/envdata.aspx) Special Administrative Regions, as well as from Taiwan province (https://taqm.epa.gov.tw) spanning from 2005 to 2020. In total, the dataset comprises 3770362 records from 1738 monitoring sites, with 1640 located in mainland China, 18 in Hong Kong, 6 in Macau, and 74 in Taiwan, as depicted in Fig. 1. In Taiwan Province, ozone concentrations measured in ppm were converted to μg/m^3 by using a factor of 1.96, following methodologies from a previous study (Yin et al. 2017). Consistent with prior ozone research (Liu et al. 2020; Zhu et al. 2022), we computed the maximum daily 8-hour average (MDA8) of ozone concentrations and selected MDA8 O_3 as the target variable for our estimation modeling. Negative values within the monitoring dataset were considered outliers and subsequently excluded. Additionally, daily MDA8 O_3 concentrations were disqualified if the valid number of hourly measurements within a natural day was less than 15 (Zhu et al. 2022). Ultimately, 184709 (4.67%) daily MDA8 O_3 records were eliminated from the monitoring dataset.

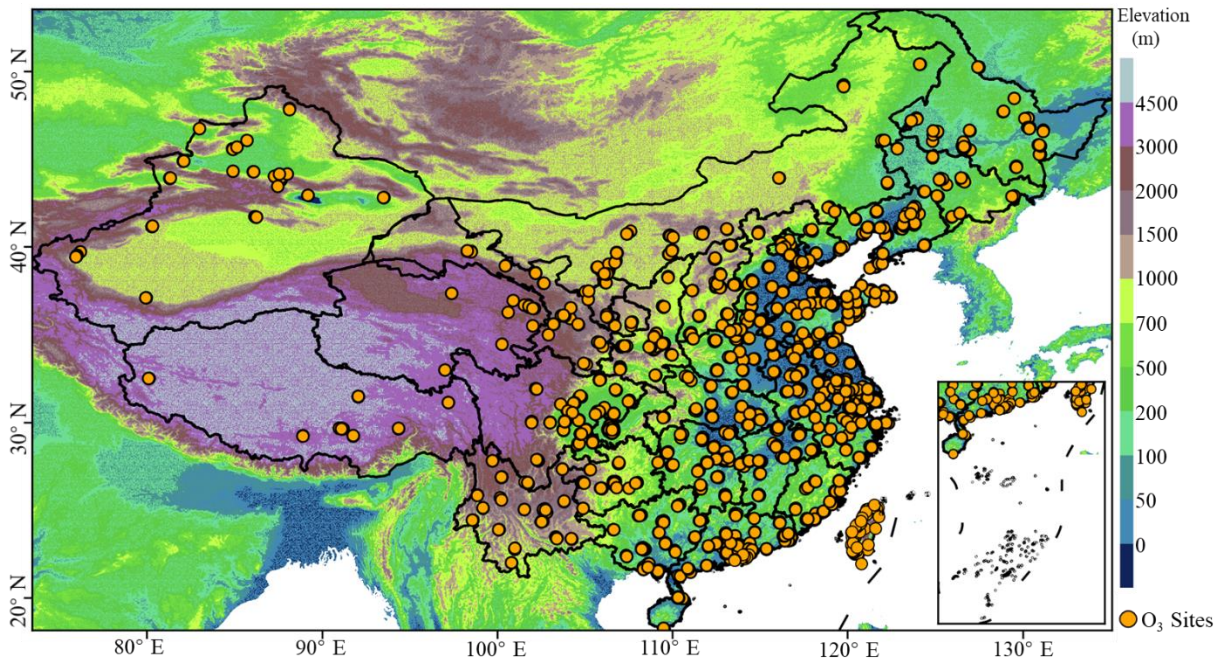

Figure S1. Geographical distribution of the study region and air quality monitoring stations, with the background of elevation.

Satellite measurements of atmospheric properties

We acquired a daily, 1-km resolution, seamless land surface temperature (LST) dataset for China covering the years 2000 to 2020 from the National Tibetan Plateau Data Center (TPDC, https://data.tpdc.ac.cn). This high-resolution dataset, referred to hereafter as TRIMS LST (Tang et al. 2024), is a product of an advanced method that merges MODIS thermal infrared observations with reanalysis datasets (Zhang et al. 2019; Zhang et al. 2021; Zhou et al. 2017). Validation against measurements from 19 surface sites confirmed its accuracy, with root mean square error (RMSE) values ranging from 0.80 to 3.68 K and mean bias error values between -2.26 and 1.73 K (Zhang et al. 2019; Zhang et al. 2021; Zhou et al. 2017). The reason for incorporating this variable is that temperature significantly influences ozone concentration by accelerating the rate of atmospheric chemical reactions, including those that produce ozone, and by increasing emissions of VOCs from biogenic sources (Sillman and Samson 1995). In contrast to previous studies that used temperature data from coarse-resolution reanalysis products, the LST dataset utilized in this study provides rich spatial gradients and captures short-term variations essential for detailed ground-level ozone analysis due to its daily, 1-km spatiotemporal resolution.

In addition, considering the correlation between ozone and particulate matter (Xue et al. 2020; Zhu et al. 2022), aerosol optical depth (AOD) was also included in the ozone modeling. We used a daily, 1-km AOD dataset covering the period from 2000 to 2020, which was developed in out previous study (He et al. 2023a). This dataset, derived from the MODIS MAIAC 1-km AOD product, was enhanced with multi-source predictors using daily random forest models. Evaluation of this dataset showed a high correlation with ground-based AOD measurements, achieving an R^2 of 0.77 and an RMSE of 0.25, which is close to the performance of the original MODIS AOD product (R^2=0.82, RMSE=0.16).

Other atmospheric parameters

We also included a range of other atmospheric parameters known to influence ground-level ozone variations in our ozone modeling. These parameters encompass meteorological elements and ozone precursors. We sourced hourly data on total precipitation (TP), U- and V-components of surface wind (U10M, V10M), surface pressure (SP), air temperature (T2M), total cloud cover (TCC), surface solar radiation downwards (SSRD), surface latent heat flux (SLHF), and planetary boundary layer height (PBLH) from the ERA5 reanalysis (https://www.ecmwf.int/). This dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), offers a global climate and weather reanalysis with a spatial resolution of 0.1°×0.1°, covering the period from 2003 to 2020. Additionally, daily surface sunshine duration (SSD) data, observed by approximately 830 monitoring stations, were obtained from the China Meteorological Data Service Center

(http://data.cma.cn/en). Hourly nitrogen dioxide (NO2) concentration data were collected from the ECMWF's fourth generation global reanalysis of atmospheric composition (EAC4, url: https://ads.atmosphere.copernicus.eu/cdsapp#!/home), which provides a spatial resolution of 0.75°x0.75° (Inness et al. 2019).

Geographical covariates

Surface-related high-resolution data measured by satellite remote sensing, including population density distribution (POP), elevation (DEM), and land-cover classification (LCC) were also collected. The Landscan annual population distribution data with 1-km spatial resolution (Rose et al. 2020) were publicly available from the Oak Ridge National Laboratory of USA (https://landscan.ornl.gov/). The 30-m elevation data was extracted from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (https://asterweb.jpl.nasa.gov/gdem.asp). The 30-m annual land cover (LC) datasets were obtained from Jie Yang and Xin Huang (Yang and Huang 2021), which is a Landsat-derived land cover product over China and contains nine classes, namely, cropland (LC1), forest (LC2), shrub (LC3), grassland (LC4), water (LC5), snow/iced (LC6), barren (LC7), impervious (LC8), and wetland (LC9). This LC dataset achieved high model performance, with overall accuracy of 79.31%, and outperforms the widely-used land cover products such as MCD12Q1 based on 5131 third-party test samples.

S1.2 Data preprocessing and integration

To facilitate the integration of variables for ground-level ozone modeling and prediction, we established a 1-km grid based on the full-coverage AOD data, resulting in a total of 9646100 grid cells across the study area. Where multiple surface ozone monitoring stations were located within the same grid cell, their readings were averaged, ultimately resulting in 3249652 samples for model training and validation. Our preprocessing techniques for handling variable datasets with differing spatial and temporal resolutions are consistent with those used in our previous studies (He et al. 2021; He et al. 2023b). Hourly atmospheric parameters of coarser resolution from ECMWF reanalysis products were first aggregated into daily averages and then downscaled to the 1-km grid using a bilinear resampling technique. Station-based surface sunshine duration (SSD) data were interpolated to the 1-km grid using inverse distance weighted interpolation. Furthermore, 30-meter land cover classification (LCC) categories were quantified as continuous values by calculating the area ratios of each land cover type within the 1-km grid cells. The details of the data sources and integration methods are provided in Table S1.

S1.3 Feature construction and selection

To account for significant temporal variations in ground-level ozone concentrations, we incorporated dummy temporal features into our model, including the day of the year (DOY) and its cosine transformation $T_x =$

$\cos\left(2\pi\frac{DOY}{365.25}\right)$. Additionally, we explored various spatial features such as latitude, longitude, Haversine distances to the four corners of the study region (Wei et al. 2023), and geospatial codes in Cartesian coordinates (Yang et al. 2022). However, preliminary analyses indicated that including these spatial features led to abnormal spatial patterns in the ozone estimates, particularly in the western areas of the study region where samples are sparse (Fig. S1). This issue of abnormal spatial patterns was also observed in particulate matter estimations from satellite remote sensing data (Ma et al. 2022b). Consequently, we decided to include only the temporal features in our modeling.

Feature selection was guided by XGBoost's impurity-based variable importance, which assesses the impact and contribution of each predictor. Eleven variables with lower importance were removed from the model, as our preliminary analyses showed that excluding these variables did not significantly affect performance (Table S2 and S3). The final set of predictors used to construct the XGBoost model included LST, SSRD, SSD, TP, AOD, NO_2, T2M, PBLH, POP, LUCT2, LUCT8, DEM, LUCT7, DOY, Tx. Further details about these variables and their abbreviations can be found in Sections 2.1.2 to 2.1.4 and Table S1.

Table S1. Data sources and preprocessing methods of variables used for ground-level ozone estimation modeling.

Variable	Abbreviation	Data source	Spatial scale	Temporal resolution	Preprocessing method
Land Surface Temperature	LST	MODIS LST: (Wan et al. 2021a, b) TPDC_LST: (Shi and Dong 2021)	1km	daily	Resampling
Sunshine duration	SSD	(NMIC 2023)	point	daily	IDW interpolation
Surface solar radiation downwards	SSRD	(Hersbach et al. 2020)	0.125°×0.125°	hourly	Resampling
2m temperature	T2M				
Total precipitation	TP				
Boundary layer height	BLH				
10 meter V wind component	V10				
10 meter U wind component	U10				
Total cloud cover	TCC				
Surface pressure	SP				
Surface latent heat flux	SLHF				
Elevation	DEM	(JAST 2019)	500m	----	Resampling
Population density	POP	(Rose et al. 2020)	1km	yearly	Area sharing
Cropland land cover	LUCT1	(Yang and Huang 2020)	30m	yearly	Area sharing
Forest land cover	LUCT2				
Shrub land cover	LUCT3				
Grassland land cover	LUCT4				
Water land cover	LUCT5				
Sonw/Ice land cover	LUCT6				
Barren land cover	LUCT7				
Impervious land cover	LUCT8				
Wetland land cover	LUCT9				
Aerosol optical depth	AOD	(He et al. 2023)	0.01°×0.01°	daily	----
Total column Nitrogen dioxide	NO2	(Inness et al. 2019)	0.75°×0.75°	hourly	Resampling
Time variables	TX	----	----	----	----
Day of year	DOY	----	----	----	----

Table S2. Variable importance ranking based on the XGBoost model with all explanatory variables.

Number	Variable	Importance of model
1	LST	0.34
2	SSRD	0.09
3	SSD	0.07
4	AOD	0.05
5	TX	0.05
6	TP	0.04
7	NO2	0.04
8	T2M	0.03
9	LUCT2	0.03
10	DOY	0.02
11	DEM	0.02
12	BLH	0.02
13	POP	0.02
14	LUCT8	0.02
15	LUCT7	0.02
16	LUCT4	0.02
17	LUCT3	0.02
18	LUCT5	0.02
19	LUCT1	0.02
20	SP	0.02
21	V10	0.02
22	TCC	0.02
23	SLHF	0.01
24	U10	0.01
25	LUCT9	0.00
26	LUCT6	0.00

Table S3. Performance comparisons of XGBoost models with various predictors.

Variables in model	R^2	RMSE (μg/m³)	MAE (μg/m³)
LST, SSD, SSRD, T2M, TP, BLH, DEM, POP, LUCT8, LUCT2, LUCT7, AOD,NO2, TX,DOY, LUCT1, LUCT3, LUCT4, LUCT5, LUCT6, LUCT9,SLHF,SP,TCC,U10,V10	0.77	21.41	15.86
LST, SSD, SSRD, T2M, TP, BLH, DEM, POP, LUCT8, LUCT2, AOD,NO2, TX,DOY	0.75	22.27	16.52
LST, SSD, SSRD, T2M, TP, BLH, DEM, POP, LUCT8, LUCT2, LUCT7, AOD,NO2, TX,DOY	0.76	22.15	16.43
LST, SSD, SSRD, T2M, TP, BLH, DEM, POP, LUCT8, LUCT2, LUCT7, AOD,NO2, TX,DOY, LUCT4	0.76	22.14	16.41

Table S4. Performance comparisons of long-term ozone estimation models over China in the literature.

Study	Spatial	Temporal		10-fold CV R^2						By-year R^2	
	Scale	Extent	Scale	Sample-based		Site-based		Site-based			
				D	M	D	M	D	M	D	M
Ma et al.(2022)	1 km	2005-2017	Daily	0.77	0.77	0.74	0.77	0.58	0.63	---	---
Liu et al.(2020)	0.1°	2005-2017	Daily	0.78	0.90	0.64	0.68	---	---	0.61	0.69
Xue et al.(2020)	0.1°	2013-2017	Daily	0.70	---	---	---	---	---	---	---
Wei et al.(2022)	10 km	2013-2020	Daily	0.87	---	0.80	---	---	---	---	---
Chen et al.(2021)	0.0625°	2008-2019	Daily	0.84	0.91	0.79	0.82	---	---	---	---
Zhu et al.(2022)	0.05°	2005-2019	Monthly	---	0.87	---	0.86	---	---	---	0.76
This study	0.01°	2000-2020	Daily	0.83	0.96	0.66	0.72	0.61	0.80	0.57	0.74

Note: D and M represent daily and monthly, respectively.

Table S5. Leave-one-year-out CV results of our proposed ozone estimation method over Hong Kong.

	Number	**R^2**	**RMSE(μg/m³)**	**MPE(μg/m³)**
Daily	23703	0.44	32.84	24.86
Monthly	1240	0.69	17.13	12.57

Table S6. Mann-Kendall test results for maximum monthly ozone concentration in China and three typical regions.

	U Statistic	P Value
China	276.00	0.04
Eastern China	278.00	0.03
PRD	204.00	0.92
NCP	271.00	0.04

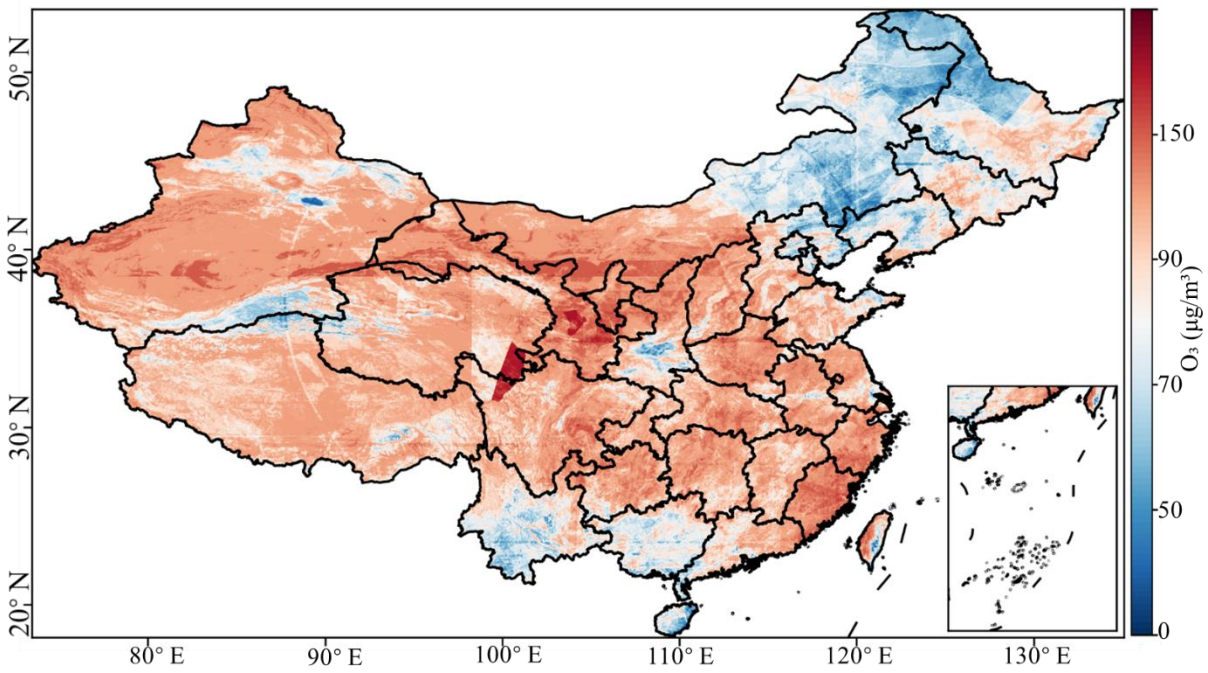

Figure S2. Abnormal spatial distribution of XGBoost predictions with additional spatial features during modeling.

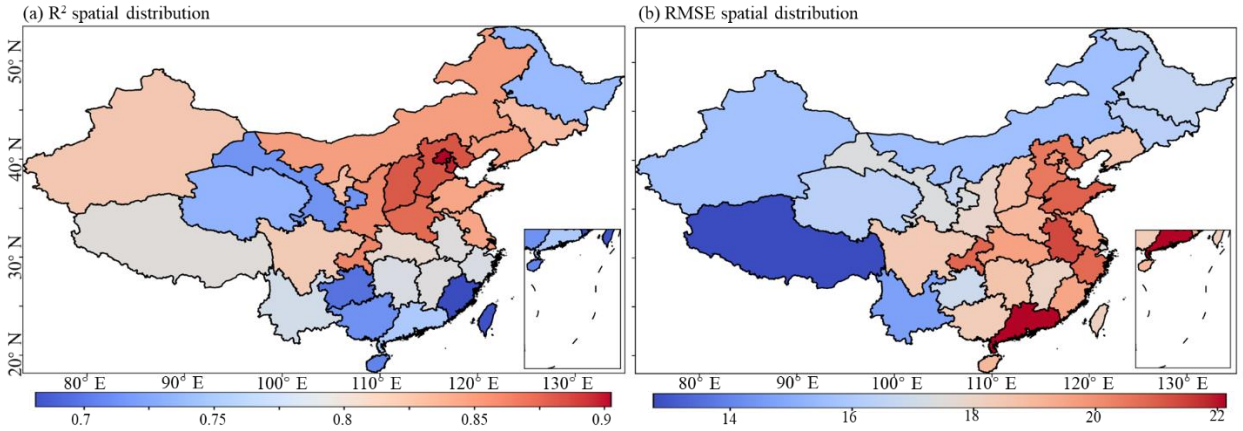

Figure S3. Spatial distribution of random 10-fold CV results of our proposed MDA8 O3 method at the provincial scale.

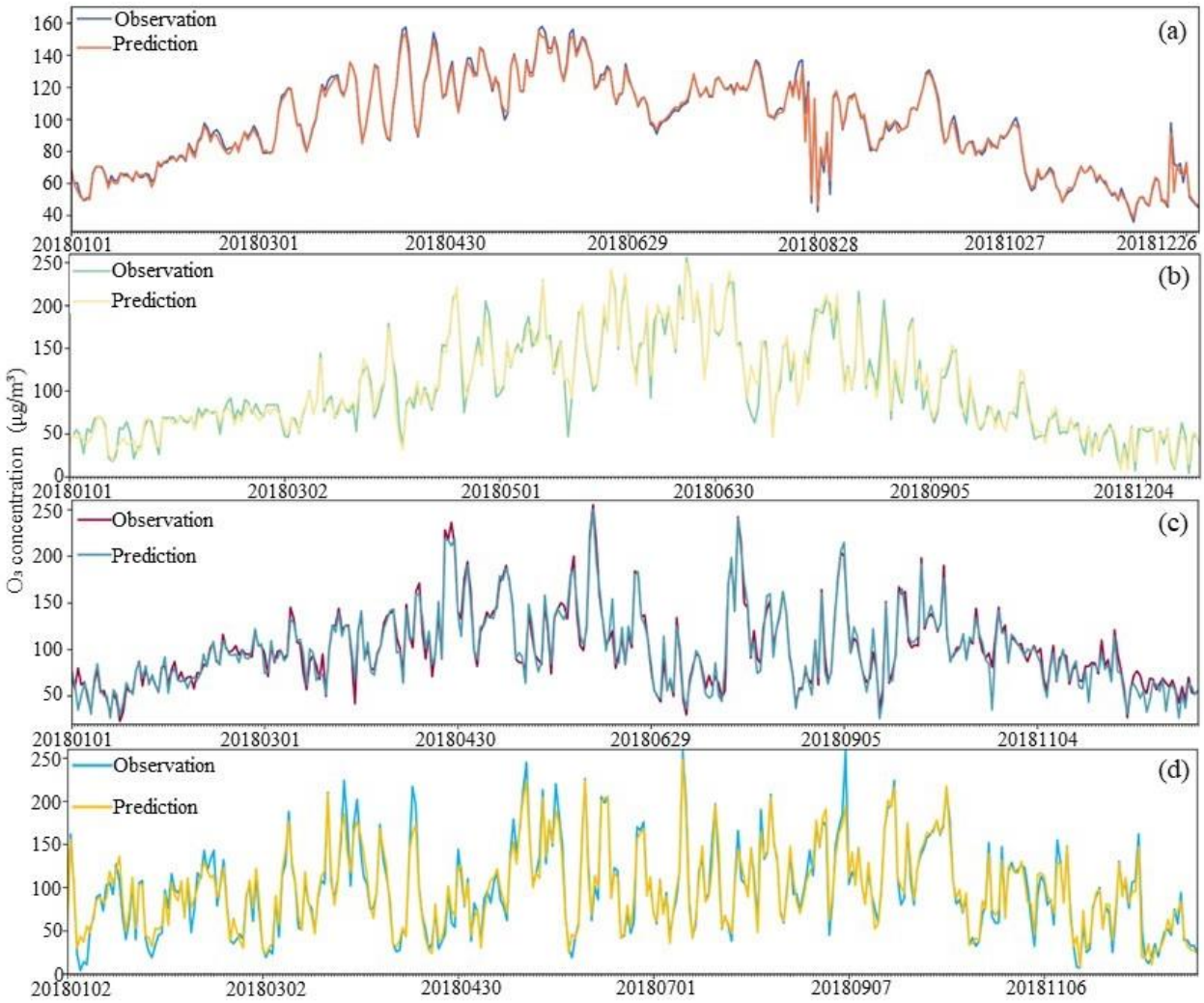

Figure S4. Time series of estimated vs. observed MDA8 O3 concentrations over China during 2018: (a) mean values at all in-situ monitors, (b) values at Wanshou Temple station in Beijing (lat=39.87°, lon=116.37°), (c) values at No.15 Factory station in Shanghai (lat= 31.20°, lon=121.48°), and (d) values at No.86 Middle School station in Guangzhou (lat= 23.11°, lon= 113.43°).

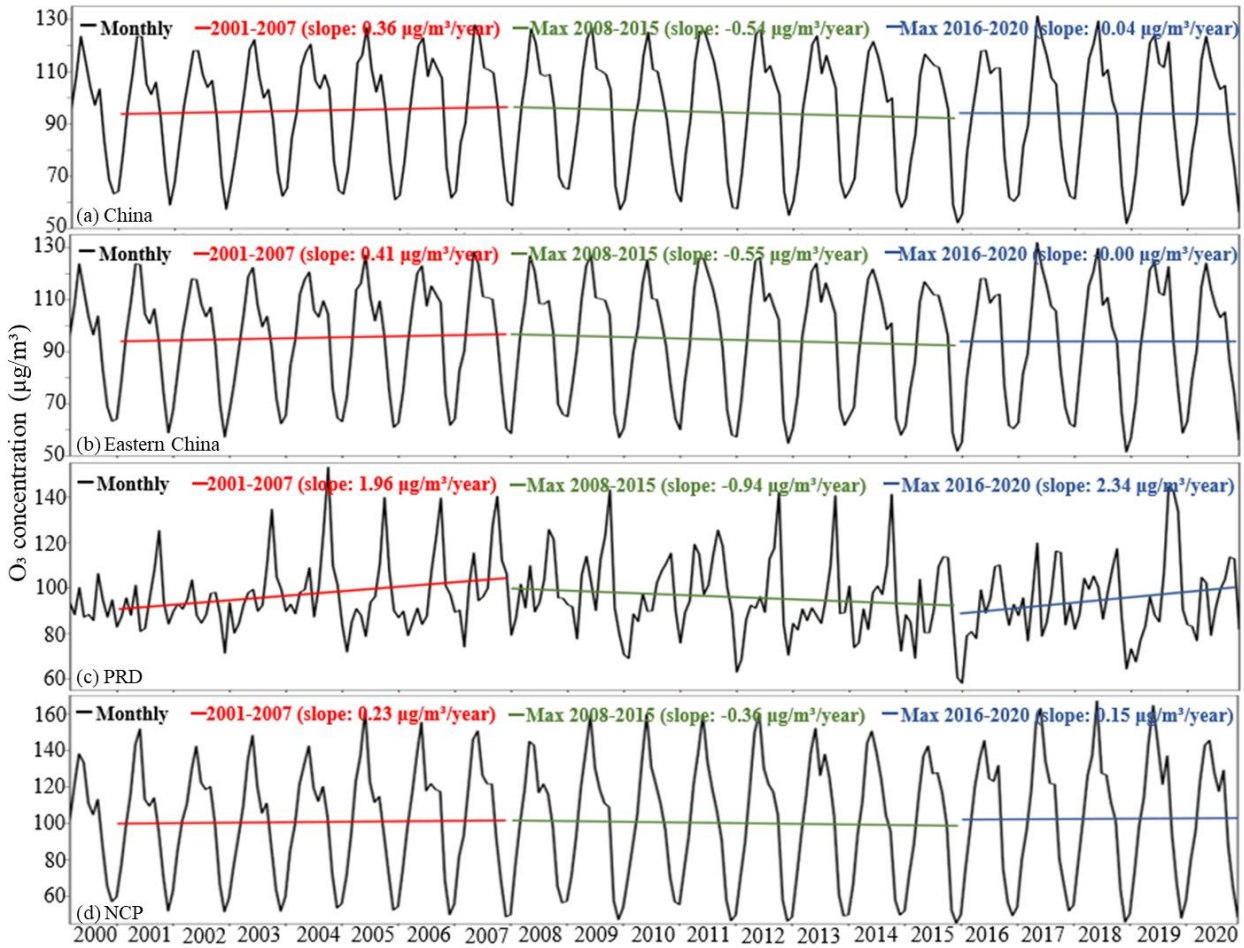

Figure S5. Time series of monthly mean population-weighted mean MDA8 O3 in China and typical exposure hotspots with linear trends.

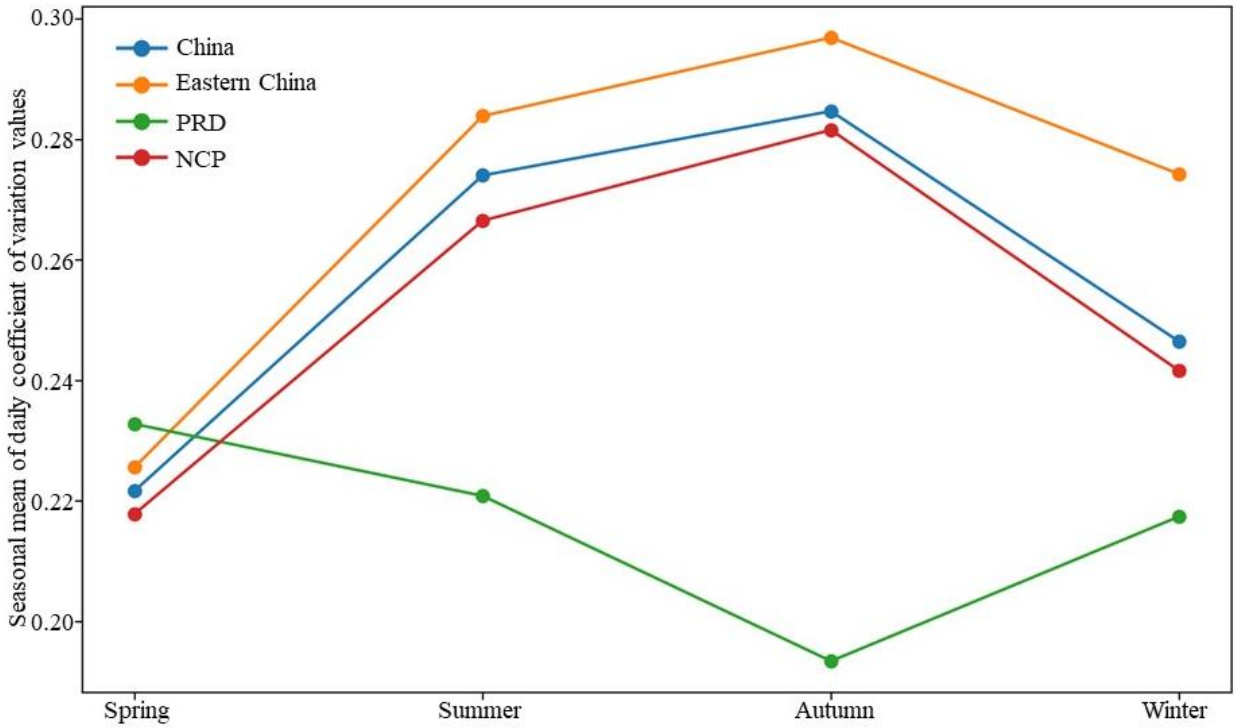

Figure S6. Seasonal mean of daily coefficient of variation values for ground-level MDA8 O3 prediction from 2000 to 2020.

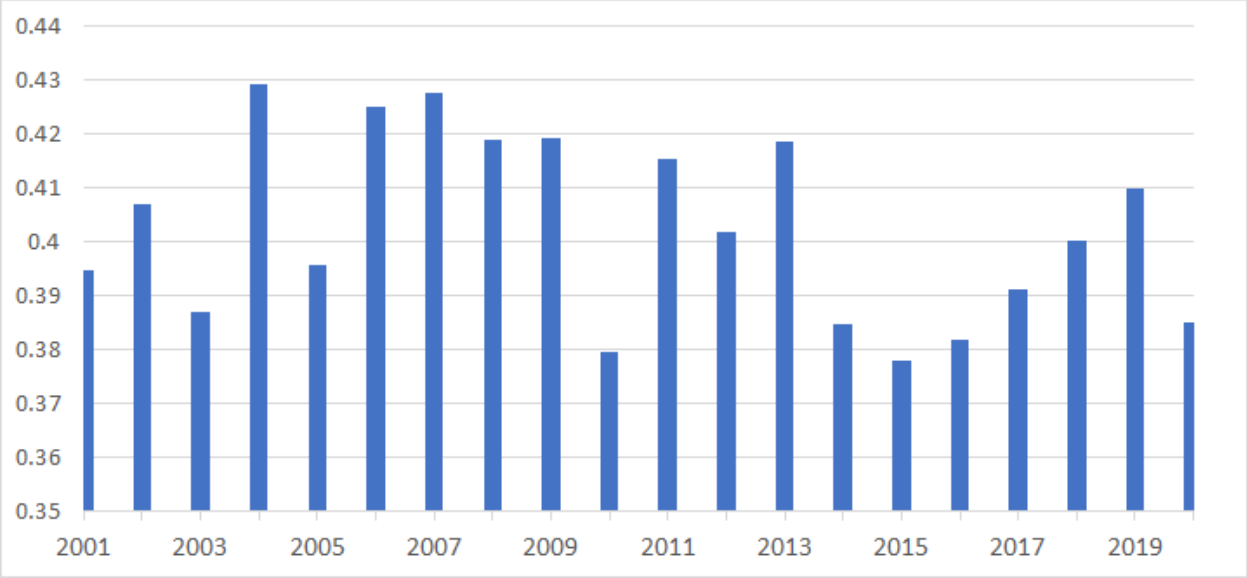

Figure S7. Average percentage of the population exposure to MDA8 O3 concentration exceeding 100 μg/m^3 over China.

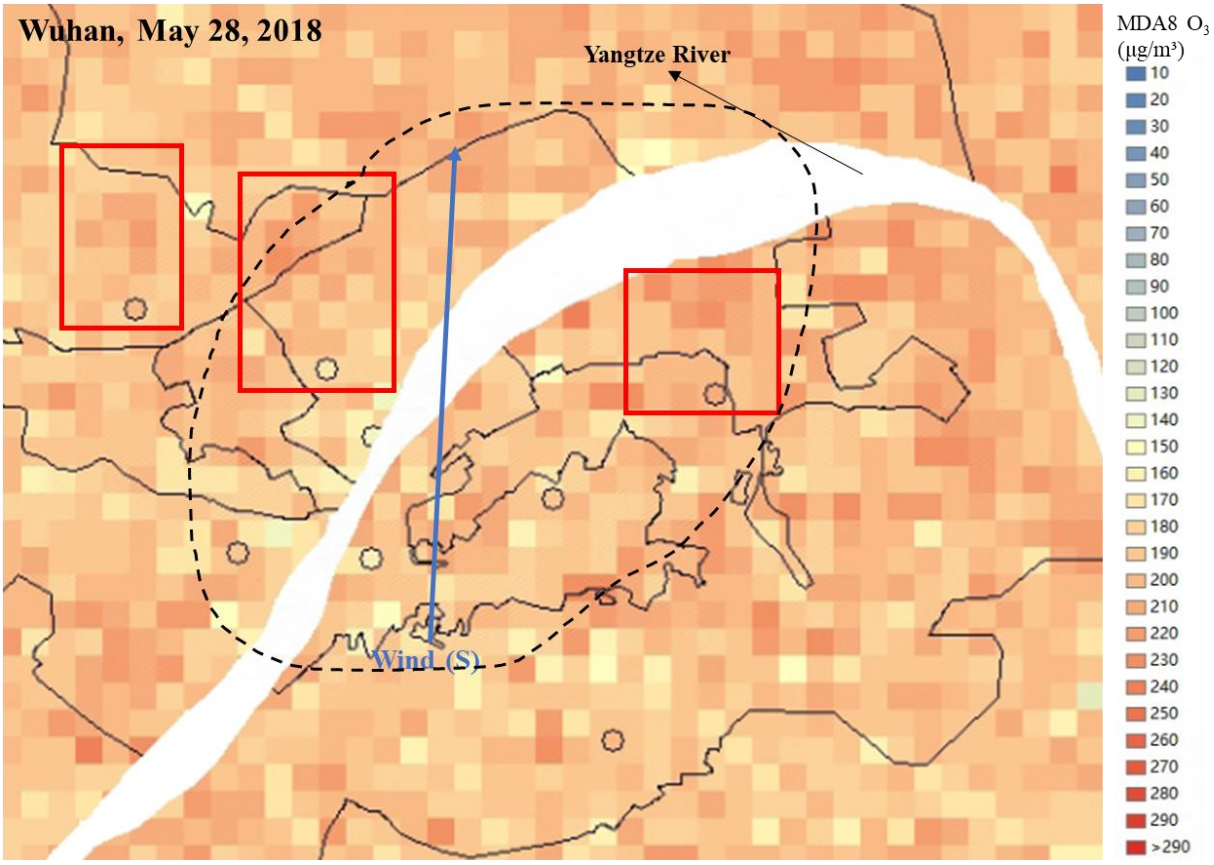

Figure S8. A case study regarding tritrated O3 from Wuhan on May 28, 2017. The downtown area of Wuhan, central to the areas surrounding the Yangtze River (typically distributed in the dashed line), is depicted in the figure. The stations are marked with dots that use the same colorbar as the MDA8 O3 concentration predictions. On this particular day, the prevailing wind direction was from the south.

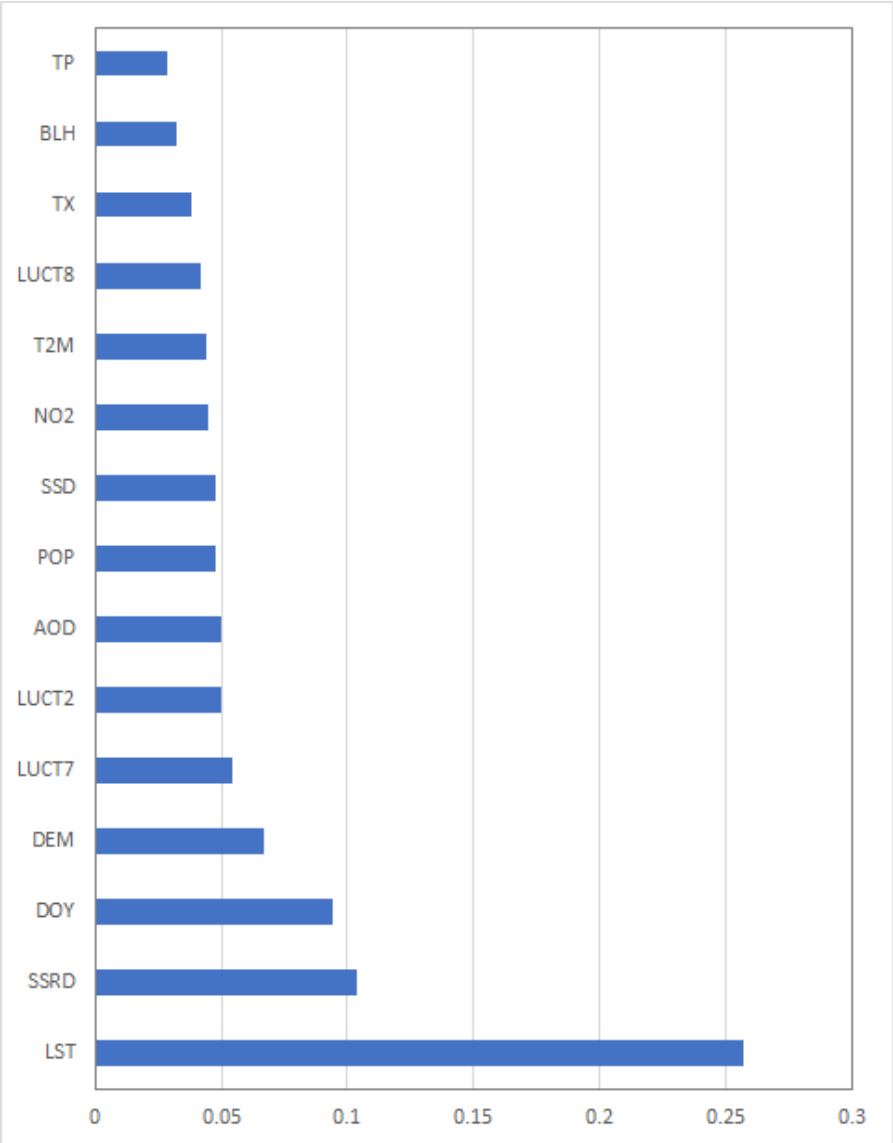

Figure S9. Importance of variables used for predicting the long-term MDA8 O3 concentration by the proposed method.

Reference

He, Q., Wang, W., Song, Y., Zhang, M., & Huang, B. (2023). Spatiotemporal high-resolution imputation modeling of aerosol optical depth for investigating its full-coverage variation in China from 2003 to 2020. *Atmospheric Research, Vol.281*, 106481(106481-106415)

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G.D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P.d., Rozum, I., Vamborg, F., Villaume, S., & Thépaut, J.N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, Vol.146*, 1999-2049

Inness, A., Ades, M., Agust, Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J.J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., & Suttie, M. (2019). The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry & Physics, Vol.19*, 3515-3556

JAST (2019). ASTER Global Digital Elevation Model V003. In: NASA EOSDIS Land Processes Distributed Active Archive Center

NMIC (2023). China Surface Meteorological Observation Data. In: National Meteorological Information Center

Rose, A., McKee, J., Sims, K., Bright, E., Reith, A., & Urban, M. (2020). LandScan Global 2019. In. Oak Ridge, TN: Oak Ridge National Laboratory

Shi, J., Cheng,Jie, & Dong, S. (2021). 1km seamless land surface temperature dataset of China (2002-2020). In C. National Tibetan Plateau Data (Ed.): National Tibetan Plateau Data, Center

Wan, Z., Hook, S., & Hulley, G. (2021a). MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061. In: NASA EOSDIS Land Processes Distributed Active Archive Center

Wan, Z., Hook, S., & Hulley, G. (2021b). MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061. In: NASA EOSDIS Land Processes Distributed Active Archive Center

Yang, J., & Huang, X. (2020). The 30 m annual land cover datasets and its dynamics in China from 1990 to 2021. In. Zenodo: Earth System Science Data, 13(1), 3907–3925,