### **Supplementary Material**

# First retrieval of 24-hourly 1-km-resolution gapless surface ozone (O<sub>3</sub>) from space in China using artificial intelligence: diurnal variations and implications for air quality and phytotoxicity

Fan Cheng<sup>1</sup>, Zhanqing Li<sup>2\*</sup>, Zeyu Yang<sup>1</sup>, Ruohan Li<sup>3</sup>, Dongdong Wang<sup>3</sup>, Aolin Jia<sup>4</sup>, Ke Li<sup>5</sup>, Bin Zhao<sup>6,7</sup>, Shuxiao Wang<sup>6,7</sup>, Dejia Yin<sup>6,7</sup>, Shengyue Li<sup>6,7</sup>,

Wenhao Xue<sup>8</sup>, Maureen Cribb<sup>2</sup>, Jing Wei<sup>2\*</sup>

- 1. College of Global Change and Earth System Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China
- 2. Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
- 3. Department of Geographical Sciences, University of Maryland, College Park, MD, USA
- 4. Department of Environment Research and Innovation, Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg
- Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, School of Environmental Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, China
- 6. State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China
- 7. State Environmental Protection Key Laboratory of Sources and Control of Air Pollution Complex, Beijing 100084, China
- 8. School of Economics, Qingdao University, Qingdao, China

\*Correspondence:

Zhanqing Li (zhanqing@umd.edu), and Jing Wei (weijing@umd.edu)

## Table of Contents

Appendix Texts	1
Appendix Figures	2
Appendix Tables	
Appendix References	

#### **Appendix Texts**

#### Text S1: Necessity of high-resolution surface O<sub>3</sub> data

The lifetime of surface O<sub>3</sub> varies considerably with location and season, ranging from a few hours in polluted urban regions up to a few weeks in the upper troposphere, as reported in many previous studies (Griffiths et al., 2021; Monks et al., 2015; Stevenson et al., 2006; Young et al., 2013). Additionally, O<sub>3</sub> exhibits a distinct chemical mechanism, notably its strong nonlinearity with NO<sub>X</sub> (Lin et al., 2018; Wang et al., 2017). For example, O<sub>3</sub> concentrations in a city center are typically lower than those in suburban areas (Xie et al., 2016), and concentrations around urban roads are generally lower than in other parts of the city center (Mavroidis and Ilia, 2012). Moreover, the substantial presence of background O<sub>3</sub> may result in smaller actual spatial disparities (Reid et al., 2008).

In our study, we examined the scale of O<sub>3</sub> variability and found the existence of strong O<sub>3</sub> heterogeneity by analyzing surface O<sub>3</sub> differences at multiple closely located sites in typical regions across China. The Kruskal-Wallis H-test and Dunn's test methods were utilized to assess the differences among these sites. When *p*-values are less than 0.05, this indicates significant differences between data from the two sites at a 95% significance level. Here, we have identified six densely populated areas with multiple closely situated ground monitoring stations across China (Figure S8) and computed the variances in surface O<sub>3</sub> levels observed at these locations in comparison to their nearby sites within the region (Figure S9). Results show that the majority of *p*-values between sites fall below 0.05, indicating pronounced spatial gradients in hourly O<sub>3</sub> concentrations among the closest sites within the  $0.1^{\circ} \times 0.1^{\circ}$  (approximately 10 km × 10 km) grid at the 95% confidence level. This underscores the importance of producing hourly O<sub>3</sub> data at a high spatial resolution (e.g., 1 km) because it facilitates the monitoring of fine-scale spatial variations in surface O<sub>3</sub> concentrations, especially in urban areas.

### **Appendix Figures**



Figure S1. Spatial distribution of ground-based surface O<sub>3</sub> monitoring stations in 2019 in China.



**Figure S2.** Comparisons of model performance (a, d, g) without, (b, e, h) with conventional, and (c, f, i) with improved spatial and temporal information for sample-CV (top row), station-CV (middle row), and block-CV (bottom row) results at 17:00 (Beijing Time) across China (sample size = 476,039). The linear-regression relation, coefficient of determination ( $\mathbb{R}^2$ ), root-mean-square error ( $\mathbb{R}MSE$ ,  $\mu g/m^3$ ), mean absolute error ( $\mathbb{M}AE$ ,  $\mu g/m^3$ ), and mean relative error ( $\mathbb{M}RE$ , %) are also given.



**Figure S3.** Out-of-station cross-validation results of hourly O<sub>3</sub> estimates ( $\mu$ g/m<sup>3</sup>) from 00:00 to 23:00 LT for 2019 in China using the 4D-STET model. Black dashed lines denote 1:1 lines, and red solid lines denote best-fit lines from linear regression. The sample size (N), linear-regression relation, coefficient of determination (R<sup>2</sup>), root-mean-square error (RMSE,  $\mu$ g/m<sup>3</sup>), mean absolute error (MAE,  $\mu$ g/m<sup>3</sup>), and mean relative error (MRE, %) are also given.



Figure S4. Out-of-block cross-validation results of hourly O<sub>3</sub> estimates (μg/m<sup>3</sup>) from 00:00 to 23:00 LT for 2019 in China using the 4D-STET model. Black dashed lines denote 1:1 lines, and red solid lines denote best-fit lines from linear regression. The sample size (N), linear-regression relation, coefficient of determination (R<sup>2</sup>), root-mean-square error (RMSE, μg/m<sup>3</sup>), mean absolute error (MAE, μg/m<sup>3</sup>), and mean relative error (MRE, %) are also given.



**Figure S5.** Surface O<sub>3</sub> concentrations (μg/m<sup>3</sup>) monitored for each hour throughout the day (00:00–23:00 LT, surrounding subplots), along with average maps during the (a) daytime (08:00–20:00 LT) and (b) nighttime (20:00–08:00 LT) in 2019 across China.



Figure S6. Boxplots of diurnal surface O<sub>3</sub> concentrations (top row; μg/m<sup>3</sup>) and the sorted SHAP importance of each variable during the daytime (08:00–20:00 LT) and nighttime (20:00–08:00 LT) in 2019 for (a) the Beijing–Tianjin–Hebei region, (b) the Yangtze River Delta region, and (c) the Pearl River Delta region. Black dashed lines represent the diurnal variations observed from ground measurements. In each box, the middle, lower, and upper horizontal black lines represent the mean bias, 25th percentile, and 75th percentile, respectively.



**Figure S7.** Boxplots of diurnal surface O<sub>3</sub> concentrations (μg/m<sup>3</sup>) covering the day for different seasons for China, the Beijing–Tianjin–Hebei (BTH) region, the Yangtze River Delta (YRD) region, and the Pearl River Delta (PRD) region. Black dashed lines represent the diurnal variations observed from ground measurements.



**Figure S8.** Selected typical densely populated areas with multiple closely located sites in China, where the number in brackets indicates the number of sites within the area.



**Figure S9.** Variances in surface O<sub>3</sub> concentrations among the sites within the six selected populated areas in China.

### **Appendix Tables**

AdaBoost     input_data n_estimators     tabular data 200       learning_rate     0.1       random_state     0       DT     input_data       min_samples_split     2       random_state     0       GBDT     input_data       n_estimators     200       alpha     0.5       max_depth     5       learning_rate     0.1       min_samples_split     2       random_state     0       OSBOT     input_data       min_samples_split     2       random_state     0       loss     'sguared error'       CatBoost     input_data       input_data     tabular data       iterations     800       depth     6       learning_rate     0.1       random_state     0       OS     squared error'       CatBoost     input_data       input_data     tabular data       witerations     2000       max_depth     5       subsimple     0.8       learning_rate     0.1       random state     0       off     random state       o     0       reg_alpha     0.1       reg_alpha     0.1 </th <th>Model</th> <th>Parameter</th> <th>Value</th>	Model	Parameter	Value
Availaboostimput_data n_estimators learning_rate random_stateadvantata out outDTinput_data min_samples_splitabular datamin_samples_split2random_state0OBDTinput_data n_estimatorsabular data abular datan_estimators2000 	AdaBoost	input data	tabular data
Inclusion200Identing rate0DTinput_datatabular datamin_samples_split2alpha0.5max_depth5learning_rate0Input_datatabular datan_estimators200alpha0.5max_depth5learning_rate0losssquared error*CatBoostinput_dataiterations800depth6learning_rate0.1min_samples_split2random_state0learning_rate0.1iterations800depth6learning_rate0.1random_state0objective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_almbda0.1random state0LightGBMinput_datanin_samples_split2random state0LightGIAM10min_child_swipt0.001learning_rate0.1reg_alpha10num_leaves250nadom state0ETinput_datarandom state0ETinput_datamin_samples_split200 <td< td=""><td>Adaboost</td><td>n estimators</td><td></td></td<>	Adaboost	n estimators	
random state0.1DTinput_datatabular datamin_samples_split2random state0GBDTinput_datatabular datan_estimators200alpha0.5max_depth5learning_rate0.1min_samples_split2random state0loss'squared error'CatBoostinput_datainput_datatabular dataiterations800depth6learning_rate0.1random state0NGBoostinput_datainput_datatabular dataiterations2000depth6learning_rate0.1random state0NGBoostinput_dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_mators200min_sitate0RFinput_datan_estimators200min_chil_samples_split2random state0LightGBMinput_datamin_chil_samples20min_chil_samples20min_chil_samples20min_site0ETinput_datarandom state0<		learning rate	0.1
Introduction     O       DT     input_data     tabular data       min_samples_split     2       GBDT     input_data     tabular data       n_estimators     200       alpha     0.5       max_depth     5       learning_rate     0.1       min_samples_split     2       random state     0       loss     'squared error'       CatBoost     input_data       iterations     800       depth     6       learning_rate     0.1       random_state     0       ZatBoost     input_data       iterations     800       depth     6       learning_rate     0.1       random state     0       XGBoost     input_data       objective     'squarederror'       rage_apha     0.1       random state     0       XGBoost     input_data       objective     'squarederror'       reg_slambda     0.1       reg_slambda <td< td=""><td></td><td>random state</td><td>0.1</td></td<>		random state	0.1
D1 mpu data min_samples_split 2 random state 0 GBDT input data ubular data 0.5 max_depth 5 learning_rate 0.1 min_samples_split 2 random_state 0 los set squared error' CatBoost input_data ubular data 0.5 max_depth 6 learning_rate 0.1 min_samples_split 0 catBoost input_data ubular data 0 depth 6 learning_rate 0.1 max_depth 6 learning_rate 0.1 max_depth 6 learning_rate 0.1 max_depth 6 learning_rate 0.1 max_depth 5 slusample 0.1 max_depth 5 slusample 0.1 max_depth 5 slusample 0.1 max_depth 0 reg_alpha 0.1 reg_alpha 0.1 regression' num_scate 0 ET input_data tabular data random state 0 ET input_data tabular data estimators 0 readom state 0 ET input_data tabular data	DT	innert dete	
InitSamples2random state0GBDTinput datatabular datan_estimators200alpha0.5max_depth5learning rate0.1min_samples_split2random_state0loss'squared_error'CatBoostinput_dataiterations800depth6learning_rate0.1random_state0loss'squared_error'XGBoostinput_datautatatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8learning_rate0.1colsample_bytree0.8learning_rate0.1colsample_bytree0.8learning_rate0.1reg_lapha0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1radom_state0lipit(datatabular dataobjective'regression'num_leaves250n_estimators2000min_samples_split2lipit(datatabular dataobjective'regression'num_leaves200min_samples_split200min_child_weight0.01learning_rate0.1l	DI	mpur_data	abular dala 2
GBDT     input_data n_estimators     200       alpha     0.5       max_depth     5       learning_rate     0.1       min_samples_split     2       random_state     0       loss     "squared error"       CatBoost     input_data       itreations     800       depth     6       learning_rate     0.1       random_state     0       loss     "squared error"       CatBoost     input_data       iterations     800       depth     6       learning_rate     0.1       random_state     0       viguarederror'     "squarederror'       nestimators     2000       max_depth     5       subsample     0.8       learning_rate     0.1       colsample_bytree     0.8       gamma     0       reg_alpha     0.1       reg_lambda     0.1       radom state     0       UiptiGBM     input_data       inpu		mm_samples_spin	2
GBD1input_data n_estimatorstabular data autorsn_estimators200alpha0.5max_depth5learning_rate0.1mom_state0los'squared error'CatBoostinput_dataiterations800depth6learning_rate0.1random_state0iterations60depth5subsample0.1random_state0viscostinput_datadepth5subsample0.8learning_rate0.1rodom state0.1objective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha0.1reg_lanbda0	CDDT		
alpha200alpha0.5max_depth5learning_rate0.1min_samples_split2random_state0loss'squared error'CatBoostinput_dataiterations800depth6learning_rate0.1random_state0objective'squared error'XGBoostinput_dataiterations800depth6learning_rate0.1r_estimators2000max_depth5subsample0.8gamma0reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_split2random state0Uiptick'regression'num_deptick250num_samples_split2random state0LightGBMinput_datainput_datatabular datanin_estimators200num_leaves250num_leaves250num_depth10min_child weight0.01learning_rate0.1bagging_fraction1.0feature_fraction0.8random state0ETinput_dataradom state0ETinput_dataradom state0andom state0andom state0andom state0 <td>GBD1</td> <td>input_data</td> <td>tabular data</td>	GBD1	input_data	tabular data
alpna0.5max_depth5learning_rate0.1min_samples_split2random_state0loss'squared error'CatBoostinput_dataiterations800depth6learning_rate0.1random state0XGBoostinput_datatiterations2000max_depth5subsample0.8cataming_rate0.1colssmple_bytree0.8gamma_rate0reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_solit2random_state0LightGBMinput_dataun_exes250n_estimators200max_depth1.0min_child_weight0.01learning_rate0.1bagging_fraction1.0reatom state0ETinput_dataunin_samples_split2random state0ETi		n_estimators	200
max_depth5leaming_rate0.1min_samples_split2random_state0loss'squared error'CatBoostinput_dataiterations800depth6learning_rate0.1random state0XGBoostinput_dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_lambda0.1reg_		aipna	0.5
learning_rate0.1min_samples_split2random_state0loss'squared error'CatBoostinput_dataiterations800depth6learning_rate0.1random_state0XGBoostinput_dataupt_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_tambda0.1reg_tambda0reg_tambd		max_depth	5
mm_samples_split2random_state0loss'squared_error'CatBoostinput_datatabular dataiterations&00depth6learning_rate0XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha10nin_samples_split200min_child_weight0.001learning_rate0.1bagging_fraction </td <td></td> <td>learning_rate</td> <td>0.1</td>		learning_rate	0.1
random_state0loss'squared_error'CatBoostinput_dataiterations800depth6learning_rate0random state0XGBoostinput_dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha10reg_alpha10reg_alpha10reg_alpha10reg_alpha1.0reg_alpha1.0reg_alpha1.0reg_alpha1.0reg_alpha1.0reg_alpha1.0reg_alpha<		min_samples_split	2
loss         'squared_error'           CatBoost         input_data iterations         tabular data           iterations         800           depth         6           learning_rate         0.1           random state         0           XGBoost         input_data         tabular data           objective         'squarederror'           n_estimators         2000           max_depth         5           subsample         0.8           learning_rate         0.1           colsample_bytree         0.8           gamma         0           reg_alpha         0.1           reg_alpha         0.1           reg_alpha         0.1           reg_alpha         0.1           reg_alpha         0.1           reg_alpha         0.1           reg_split         2           random_state         0           LightGBM         input_data           n_estimators         200           min_samples_split         2           n_deta         10           min_chil_samples         200           max_depth         10           min_chil_samples		random_state	0
CatBoostinput_datatabular dataiterations800idepth6learning_rate0.1random state0XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alond0.1reg_alpha0.1reg_bandom200min_samples_split2random state0LightGBMinput_datanum_leaves2000max_depth10min_samples_split2random state0LightGBMinput_datanum_leaves2000max_depth10min_samples_split200max_depth10min_child_weight0.001learning_rate0.001learning_rate0.001learning_rate0.001feature_fraction0.8random state0ETinput_dataestimators200min_samples_split20ingut_data0.8random_state0ETinput_dataestimators200min_samples_split20min_samples_split20min_samples_split20<		loss	'squared_error'
iterations800depth6learning_rate0.1random state0XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lanbda0.1reg_alpha0.1reg_lanbda0.1random_state0RFinput_datainput_data200min_samples_split2random_state0LightGBMinput_dataobjective'regression'num_leaves250n_estimators2000max_depth10min_samples_split2random state0LightGBMinput_datamin_child_samples20max_depth10min_child_weight0.001learning_rate0.1bagging_fraction0.8random_state0ETinput_datainput_data0random_state0ETinput_datainput_data10random_state0ETinput_datarandom_state0ETinput_datarandom_state0Iterrity_random state0Iterrity_random state0Iterrity_random state0Iterrity_ra	CatBoost	input_data	tabular data
depth6learning_rate0.1random_state0XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0.1reg_alpha0reg_alpha0n_estimators200min_samples_split2random_state0LightGBMinput_datanum_leaves250n_estimators2000min_child_samples2000max_depth10min_child_samples200min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8min_child_weight0.001learning_rate0.8min_child_samples0.8min_samples_split20feature_fraction0.8min_samples_split0.8random_state0ETinput_datamin_samples_split200min_samples_split200min_samples_split0.8random_state0estimators200min_samples_split200random_state0estimato		iterations	800
learning_rate0.1random_state0XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_low2000reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_stimators200min_samples_split2random state0lightGBMinput_dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state001min_samples_split2random_state001min_samples_split2random_state0 </td <td></td> <td>depth</td> <td>6</td>		depth	6
random state0XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_alpha0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random state0LightGBMinput_datan_estimators200max_depth10min_samples_split2random state0LightGBMinput_datan_estimators2000max_depth10min_samples_split250n_estimators2000max_depth10min_child_samples20min_child_samples20min_child_samples20min_child_samples0ETinput_datarandom_state0ETinput_datamin_samples_split2min_samples_split2000min_samples_split2000max_depth0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0uim_andem_split2min_samples_split2min_samples_split2min_samples_split2min_samples_split2 </td <td></td> <td>learning_rate</td> <td>0.1</td>		learning_rate	0.1
XGBoostinput_datatabular dataobjective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_alpha0reg_sample_systee0RFinput_datan_estimators200min_samples_split2random_state0LightGBMinput_datan_estimators200n_estimators200min_samples_split2random_state0LightGBMinput_datan_estimators200n_estimators200num_leaves250n_estimators200min_child_samples20min_child_samples20min_child_samples0.1ETinput_dataeatimators0.001ETinput_datainput_data0.8random_state0ETinput_dataestimators200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200		random_state	0
objective'squarederror'n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1random_state0RFinput_data200min_samples_split2random state0LightGBMinput_datatabular dataobjective'regression'num_leaves250ne_stimators2000max_depth10min_child_samples2000max_depth0.011min_child_samples0.001mar_depth0.001mar_faction0.1min_state0ETinput_dataestimators200min_state0ETinput_datamin_samples_split200andom state0ETinput_dataestimators200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split <td>XGBoost</td> <td>input_data</td> <td>tabular data</td>	XGBoost	input_data	tabular data
n_estimators2000max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1random_state0RFinput_data n_estimators200min_samples_split2random_state0LightGBMinput_data objective'regression' regression' num_leavesn_estimators2000max_depth10min_child_samples200min_child_samples200max_depth10min_child_weight0.01learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datamin_shild_samples20min_child_weight0.01learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datamin_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split <td></td> <td>objective</td> <td>'squarederror'</td>		objective	'squarederror'
max_depth5subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1reg_lambda0.1reg_lambda0nestimators200min_samples_split2random_state0LightGBMinput_datanput_datatabular dataobjective'regression'n_estimators2000min_samples_split2random_state0UightGBMinput_dataobjective'regression'n_estimators2000max_depth10min_child_samples20n_astimators2000max_depth10min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datastimators200min_samples_split2andom_state0objectives200min_samples_split200andom_state200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min		n estimators	2000
subsample0.8learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random_state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_samples200min_child_samples0.11bagging_fraction1.0feature_fraction0.01eatmators0.001ETinput_dataestimators0ETinput_dataestimators200min_state0IETinput_dataestimators200random_state0IETinput_dataestimators200min_samples_split2random_state0IETinput_dataestimators200min_samples_split200random_state0IETinput_dataestimators200min_samples_split2random_state0random_state0random_state0random_state0random_state0random_state0ran		max depth	5
learning_rate0.1colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random_state0LightGBMinput_dataobjective'regression'num_leaves250n_estimators2000max_depth10min_chil_samples20min_chil_samples20min_chil_samples0.1learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataatbular data0estimators200min_chil_samples20min_chil_samples20min_chil_samples20min_chil_samples20min_chil_samples20min_chil_samples20min_chil_samples20min_chil_samples20min_state0ETinput_dataestimators200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split200min_samples_split </td <td></td> <td>subsample</td> <td>0.8</td>		subsample	0.8
colsample_bytree0.8gamma0reg_alpha0.1reg_lambda0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random_state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datamax_depth0ETinput_datamax_depth0ETinput_dataestimators200max_depth0num_leaves20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2200200		learning rate	0.1
gamma0reg_alpha0.1reg_lambda0.1reg_lambda0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random_state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datatabular data200min_samples_split22001.0feature_fraction0.8random_state0ETinput_datatabular data200min_samples_split2200200min_samples_split2random_state0		colsample bytree	0.8
reg_alpha0.1reg_lambda0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random_state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datamin_samples_split200min_samples_split2.0min_state0.1max_depth1.0min_child_weight0.001eature_fraction0.8min_state0ETinput_datamin_samples_split2.00min_samples_split2.00min_samples_split0.1min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split2.00min_samples_split <td< td=""><td></td><td>gamma</td><td>0</td></td<>		gamma	0
reg_lambda0.1random_state0RFinput_datatabular datan_estimators200min_samples_split2random_state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0random_state0ETinput_datatabular dataestimators200min_child_weight0.8random_state0ETinput_datatabular dataestimators200min_samples_split2random_state0ETinput_dataestimators200min_samples_split2random_state0input_datatabular dataestimators200min_samples_split2random_state0input_datatabular dataestimators200min_samples_split2random_state0input_data10random_state0input_data10random_state0input_data10random_state0input_data10input_data10input_data10input_data10input_data10 <td></td> <td>reg alpha</td> <td>0.1</td>		reg alpha	0.1
random state0RFinput_datatabular datan_estimators200min_samples_split2random state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_datamin_samples_split2000max_depth0.1learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators2000min_samples_split2andom_state0ETinput_dataestimators200min_samples_split2random_state0		reg lambda	0.1
RFinput_data n_estimatorstabular data 200min_samples_split random state2random state0LightGBMinput_data objectivetabular data 'regression'num_leaves n_estimators250n_estimators2000max_depth10min_child_samples paging_fraction feature_fraction0.1bagging_fraction random_state1.0ETinput_data random_state0ETinput_data random_state2000min_samples_split random_state20c1.0c20random_state0C200random_state0ETinput_data estimators random_state2000random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_state0C200random_st		random state	0
Image: minestimators minestimators200minestimators2random state0LightGBMinput_datatabular dataobjective num_leaves'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction feature_fraction random_state1.0ETinput_dataabular dataestimators200max_adepth0component2.0min_child_weight0.001learning_rate0.1bagging_fraction feature_fraction random_state2.0ETinput_data estimators2.00min_samples_split random_state2.00Imput_data2.00min_samples_split random_state2.00Imput_data2.00I	RF	input data	tabular data
min_samples_split random_state2LightGBMinput_data objectivetabular dataobjective num_leaves'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction random_state1.0ETinput_data estimators0ETinput_data estimators2000min_samples_split random_state20001.01200011.01200011.0120011.011.0120011.01 <td></td> <td>n estimators</td> <td>200</td>		n estimators	200
random state0LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2child_asamples0child_state0		min samples split	2
LightGBMinput_datatabular dataobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2constant200min_state0		random state	$\frac{1}{0}$
Lighter billimpla_interimpla_interobjective'regression'num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0	LightGBM	input data	tabular data
num_leaves250n_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0	LightODM	objective	'regression'
n_estimators200max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0		num leaves	250
In_estimators2000max_depth10min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0		n estimators	2000
min_child_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0		max_denth	10
min_cinid_samples20min_child_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0		min_child_samples	20
Inim_cinitd_weight0.001learning_rate0.1bagging_fraction1.0feature_fraction0.8random_state0ETinput_dataestimators200min_samples_split2random_state0		min_child_weight	0.001
bagging_fraction       1.0         feature_fraction       0.8         random_state       0         ET       input_data         estimators       200         min_samples_split       2         random_state       0		learning rate	0.1
bagging_inaction       1.0         feature_fraction       0.8         random_state       0         ET       input_data       tabular data         estimators       200         min_samples_split       2         random_state       0		hagging fraction	0.1
random_state     0.8       ET     input_data     tabular data       estimators     200       min_samples_split     2       random_state     0		facture fraction	1.0
ET     input_data     tabular data       estimators     200       min_samples_split     2       random state     0		readure_inaction	0.8
E 1 input_data tabular data estimators 200 min_samples_split 2 random state 0	ET		U tahulan 1 t
estimators     200       min_samples_split     2       random state     0	EI	input_data	tabular data
min_samples_split     2       random state     0		estimators	200
random_state 0		min_samples_split	2
		random_state	0

 Table S1. Summary of the main parameters of tree-based machine-learning models.

DF	input_data	tabular data
	n_estimators	2
	n_trees	100
	max_layers	20
	criterion	'squared_error'
	predictor	'xgboost'
	random_state	0

Model	Parameter	Value
MLP	input data	tabular data
	input size	25
	hidden layers	1
	hidden size	64
	data normalization	'StandardScaler'
	train epochs	50
	batch size	64
	activation function	Relu()
	output activation	Linear()
	learning rate	0.001
	optimizer	Adam()
	loss function	MAELoss()
DBN	input data	tabular data
DDI	input size	25
	RBM num	25
	hidden sizes	$\frac{2}{30}$ 10
	data normalization	'StandardScaler'
	train enochs	100
	hatch size	64
	activation function	Sigmoid()
	output activation	L inear()
	learning rate	
	ontimizer	
	loss function	MAEL oss()
PasNat	input data	tabular data
Keshet	input_data	
	Residual Block num	25
	data normalization	2 'StandardScaler'
	train enochs	50
	hatah siza	50
	activation function	Delu()
	autivation function	L inear()
	loarning rate	
	ontimizer	0.001 Adam()
	loss function	MAEL oss()
CNINI		
CNN	input_data	tabular data
	input channel	1
	input size	25 (1, 2)
	kernel size	(1, 3)
		3
	hidden size	16, 32, 64
	data normalization	'StandardScaler'
	train epochs	50
	batch size	64 D 1 ()
	activation function	Relu()
	output activation	Linear()
	learning rate	0.001
	optimizer	Adam()
	loss function	MAELoss()
LSTM	input_data	tabular data
	input size	25
	hidden layers	3
	hidden size	90

Table S2. Summary of the main parameters of deep-learning models used in this study.

	data normalization	'StandardScaler'
	train epochs	150
	batch size	256
	output activation	Linear()
	learning rate	0.001
	optimizer	Adam()
	loss function	MAELoss()
ResNeXt	input_data	tabular data
	input size	25
	cardinality	32
	Residual Layer num	4
	Residual Block num	[3, 4, 6, 3]
	data normalization	'StandardScaler'
	train epochs	50
	batch size	64
	activation function	Relu()
	output activation	Linear()
	learning rate	0.001
	optimizer	Adam()
	loss function	MAELoss()

17:00 LT in China based on the station-CV approach.							
Category	Core model	R <sup>2</sup>	Slope	RMSE	MAE	Speed (s)	Memory (GB)
Tree-based	AdaBoost	0.492	0.451	38.94	30.80	504.45	0.1068
machine	DT	0.725	0.855	30.12	20.22	12.16	0.0319
learning	GBDT	0.753	0.739	26.86	20.23	1000.68	0.1045
	CatBoost	0.783	0.769	25.19	18.96	21.24	0.1482
	XGBoost	0.815	0.803	23.27	17.40	247.53	0.1314
	RF	0.863	0.826	20.13	14.45	324.62	6.0102
	LightGBM	0.873	0.857	19.29	14.19	47.18	0.1495
	ET	0.875	0.792	20.36	14.64	64.81	10.3010
	DF	0.888	0.888	18.08	12.86	9491.03	11.3705
Deep	DBN	0.716	0.735	28.85	21.90	156.81	0.0624
learning	MLP	0.724	0.705	28.47	21.62	175.11	0.0930
	CNN	0.777	0.783	25.79	19.40	1282.64	0.1698
	LSTM	0.784	0.796	25.14	18.81	3719.45	0.1556
	ResNet	0.799	0.820	24.23	18.09	2628.15	1.0659
	ResNeXt	0.820	0.835	22.94	17.01	8144.82	1.1123

**Table S3.** Performance and efficiency comparisons of different 4-dimensional space-time (4D-ST) tree-based machine- and deep-learning models for estimating surface O<sub>3</sub> concentrations at 17:00 LT in China based on the station-CV approach.

AdaBoost: Adaptive Boosting; DT: Decision Trees; GBDT: Gradient Boosting Decision Tree; CatBoost: Categorical Boosting; XGBoost: eXtreme Gradient Boosting; RF: Random Forest; LightGBM: Light Gradient Boosting Machine; ET: ExtraTrees; DF: Deep Forest; MLP: Multilayer Perceptron; DBN: Deep Belief Network; ResNet: Deep Residual Network; CNN: Convolutional Neural Network; LSTM: Long Short Term Memory; ResNeXt: ResNet Next.

1/:00 LT in China based on the block-CV approach.							
Category	Core model	R <sup>2</sup>	Slope	RMSE	MAE	Speed (s)	Memory (GB)
Tree-based	DT	0.555	0.749	39.55	29.06	12.78	0.0366
Machine	AdaBoost	0.572	0.494	39.94	31.60	359.81	0.0106
Learning	XGBoost	0.745	0.729	27.88	20.90	403.70	0.1265
	GBDT	0.747	0.716	28.55	21.01	894.83	0.1056
	RF	0.763	0.729	27.62	19.95	335.28	5.8400
	LightGBM	0.765	0.750	26.75	20.01	48.75	0.2034
	CatBoost	0.771	0.757	26.41	19.71	14.22	0.1939
	ET	0.774	0.737	26.34	19.48	67.44	7.4371
	DF	0.775	0.751	26.87	19.33	9914.96	10.6786
Deep	MLP	0.714	0.694	30.33	22.69	365.61	0.0934
Learning	DBN	0.731	0.716	29.50	21.82	628.19	1.9854
	ResNet	0.737	0.733	29.06	21.11	2453.79	1.0668
	ResNeXt	0.747	0.747	28.48	20.68	7165.49	1.2687
	LSTM	0.752	0.731	28.25	20.60	4367.56	0.1539
	CNN	0.754	0.770	28.12	20.38	3681.92	1.1686

**Table S4.** Performance and efficiency comparisons of different 4-dimensional space-time (4D-ST) tree-based machine- and deep-learning models for estimating surface O<sub>3</sub> concentrations at 17:00 LT in China based on the block-CV approach.

AdaBoost: Adaptive Boosting; DT: Decision Trees; GBDT: Gradient Boosting Decision Tree; CatBoost: Categorical Boosting; XGBoost: eXtreme Gradient Boosting; RF: Random Forest; LightGBM: Light Gradient Boosting Machine; ET: ExtraTrees; DF: Deep Forest; MLP: Multilayer Perceptron; DBN: Deep Belief Network; ResNet: Deep Residual Network; CNN: Convolutional Neural Network; LSTM: Long Short Term Memory; ResNeXt: ResNet Next.

### **Appendix References**

- Griffiths, P.T., Murray, L.T., Zeng, G., Shin, Y.M., Abraham, N.L., Archibald, A.T., Deushi, M., Emmons, L.K., Galbally, I.E., Hassler, B., Horowitz, L.W., Keeble, J., Liu, J., Moeini, O., Naik, V., O'Connor, F.M., Oshima, N., Tarasick, D., Tilmes, S., Turnock, S.T., Wild, O., Young, P.J., Zanis, P., 2021. Tropospheric ozone in CMIP6 simulations. Atmos. Chem. Phys. 21, 4187–4218. https://doi.org/10.5194/acp-21-4187-2021
- Lin, Y., Zou, J., Yang, W., Li, C.-Q., 2018. A review of recent advances in research on PM<sub>2.5</sub> in China. Int. J. Environ. Res. Public Health 15, 438. https://doi.org/10.3390/ijerph15030438
- Mavroidis, I., Ilia, M., 2012. Trends of NO<sub>x</sub>, NO<sub>2</sub>, and O<sub>3</sub> concentrations at three different types of air quality monitoring stations in Athens, Greece. Atmos. Environ. 63, 135–147. https://doi.org/10.1016/j.atmosenv.2012.09.030
- Monks, P.S., Archibald, A.T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K.S., Mills, G.E., Stevenson, D.S., Tarasova, O., Thouret, V., Von Schneidemesser, E., Sommariva, R., Wild, O., Williams, M.L., 2015. Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. Atmos. Chem. Phys. 15, 8889–8973. https://doi.org/10.5194/acp-15-8889-2015
- Reid, N., Yap, D., Bloxam, R., 2008. The potential role of background ozone on current and emerging air issues: An overview. Air Qual. Atmos. Health 1, 19–29. https://doi.org/10.1007/s11869-008-0005-z
- Stevenson, D.S., Dentener, F.J., Schultz, M.G., Ellingsen, K., Noije, T.P.C. van, Wild, O., Zeng, G., Amann, M., Atherton, C.S., Bell, N., Bergmann, D.J., Bey, I., Butler, T., Cofala, J., Collins, W.J., Derwent, R.G., Doherty, R.M., Drevet, J., Eskes, H.J., Fiore, A.M., Gauss, M., Hauglustaine, D.A., Horowitz, L.W., Isaksen, I.S.A., Krol, M.C., Lamarque, J.-F., Lawrence, M.G., Montanaro, V., Müller, J.-F., Pitari, G., Prather, M.J., Pyle, J.A., Rast, S., Rodriguez, J.M., Sanderson, M.G., Savage, N.H., Shindell, D.T., Strahan, S.E., Sudo, K., Szopa, S., 2006. Multimodel ensemble simulations of present-day and near-future tropospheric ozone. J. Geophys. Res. Atmos. 111. https://doi.org/10.1029/2005JD006338
- Wang, T., Xue, L., Brimblecombe, P., Lam, Y.F., Li, L., Zhang, L., 2017. Ozone pollution in China: a review of concentrations, meteorological influences, chemical precursors, and effects. Sci. Total Environ. 575, 1582–1596. https://doi.org/10.1016/j.scitotenv.2016.10.081
- Xie, M., Zhu, K., Wang, T., Chen, P., Han, Y., Li, S., Zhuang, B., Shu, L., 2016. Temporal characterization and regional contribution to O<sub>3</sub> and NO<sub>x</sub> at an urban and a suburban site in Nanjing, China. Sci. Total Environ. 551–552, 533–545. https://doi.org/10.1016/j.scitotenv.2016.02.047
- Young, P.J., Archibald, A.T., Bowman, K.W., Lamarque, J.-F., Naik, V., Stevenson, D.S., Tilmes, S., Voulgarakis, A., Wild, O., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W.J., Dalsøren, S.B., Doherty, R.M., Eyring, V., Faluvegi, G., Horowitz, L.W., Josse, B., Lee, Y.H., MacKenzie, I.A., Nagashima, T., Plummer, D.A., Righi, M., Rumbold, S.T., Skeie, R.B., Shindell, D.T., Strode, S.A., Sudo, K., Szopa, S., Zeng, G., 2013. Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP). Atmos. Chem. Phys. 13, 2063–2090. https://doi.org/10.5194/acp-13-2063-2013