Reviewer 1#

This is an interesting study, using machine learning to estimate the ozone formation sensitivity. The idea is not novel (a few previous studies with similar scope are cited in this manuscript). The method, using reactivity-corrected VOC measurements (i.e., initial VOC concentrations), sheds some insights into ozone production in an urban environment.

Reply: Thanks for your positive comments. We have carefully responded to all of your **point-by-point** comments and issues and have revised the manuscript accordingly. These revisions are described in detail below.

However, there are several major issues:

Q1: (1) The machine learning workflow described in this manuscript does not include a robust or systematic solution to mitigate overtraining. I will elaborate on this later but the measures described in this work absolutely do not guarantee that overtraining is/can be avoided.

Reply: Thank you for your good suggestion. According to your suggestion, we performed a 12-fold cross-validation after data-normalization, i.e., by randomly dividing the dataset into 12 subsets and alternately taking one subset as testing data and the rest as training data. By doing this, every data point has an equal chance being trained and tested. In lines 148-153 in the revised manuscript, we added a short paragraph "To avoid over-fitting, we trained the random forest model using cross-validation for the normalized data, which can improve the robustness of the model. Briefly, we randomly divided the normalized data into 12 subsets, then alternately took one subset as testing data along with the rest as training data. By doing this, every data point has an equal chance being trained and tested.".

We added the RF model workflow to Text S3 in the revised Supporting Information.

"**Text S3. Workflow of RF model and the calculation of Relative Importance (RI)**

The workflow of RF model used in this study was established through the following steps.

(1) Data description. The length of the input data from 2014 to 2016 were 1190, 1062 and 872 rows, respectively, in which different types of VOCs, NOx, CO, $PM_{2.5}$ and meteorological parameters (including temperature, relative humidity, solar radiation, wind speed and direction) were used as input variables and O_3 was the output variable. The mean values (±standard deviation) of input/output parameters are shown in Table S1.

(2) Data process. After the extreme values were removed, all data were normalized (between 0 and 1) in order to decrease the sample distribution range, accelerate calculation efficiency and improve the robustness of the RF model. Then, the dataset was randomly divided into 12 subsets. Thus, a 12-fold cross-validation was performed by alternately taking one subset as testing data and the rest as training data to ensure that each data point has an equal chance being trained and tested.

(3) Hyper-parameters optimization. All network configuration parameters (i.e., leaf number, number of trees, algorithm, and so on) were modified by a trial and error method to obtain the optimized network structure. The optimized RF model parameters are shown in Table S2. Figures S13 and S14 show the examples to optimize the number of minimal samples split and trees, respectively.

(4) Model uncertainty estimation. The uncertainty of the model was estimated according to the predicted and observed O_3 concentrations. The performance of the model was evaluated using R square (R^2) and Root Mean Squared Error (RMSE).

(5) Relative importance (RI) calculation: The influence of an input variable on model performance was evaluated by changes in the accuracy of the model by variable permuting. Briefly, a change of prediction error was resulted from permuting a variable across the observations. The magnitude of the response was estimated using out-of-bag error of a predictor according to following steps.

For a random forest model that has T learners and p predictors in the training data, the first step is to identify the out-of-bag observations and the indices of the predictor variables that are split to a growing tree t (from 1 to T). Then, one can estimate the out-of-bag error (ε_t) for each tree. For a predictor variable x_j (j: from 1 to p), one can estimate the model error ($\varepsilon_{t,\,j}$) again corroding to the out-of-bag observations after randomly

permuting the observations of x_j. Thus, the difference of the model error ($d_{t,j} = \varepsilon_{t,j} - \varepsilon_t$) is obtained. If the predictor variables are not split, the difference of a growing tree t is 0. The second step is to calculate the mean difference of the model errors ($\overline{d}_j$), and the standard deviation (σ_j) of the differences for all the learners and each predictor variable in the training data. Finally, the out-of-bag relative importance (RI) for x_j is calculated by dividing the difference of the model errors by the standard deviation ($\overline{d}_j/\sigma_j$).

(6) EKMA curves. The Empirical Kinetic Modeling Approach (EKMA) was used to assess the O_3 formation mechanism regime. Both the RF model and a box model with Master Chemical Mechanism (MCM, 3.3.1) were used to calculate the EKMA curves. For the RF model simulations, the observed point data was chosen as the mean values of the input parameters during our observations, then the concentrations of VOCs and NOx were varied ±10% (or from 90% to 110%) of their mean values with a step of 1% in a two-dimensional matrix along with other inputs unchanged. This matrix was used as the testing data, while all the measured data were taken as the training data in the RF model to simulate O_3 concentrations under different scenarios of VOCs and NOx concentrations. To decrease the model uncertainty, we set relatively small variations of VOCs and NOx (±10%) compared to the observed values in this study. The mean relative error of simulated O_3 concentrations between RF model and Box model (within 15.6%, Figure S8) suggests that the RF model can well predict O_3 concentrations during our observations."

Q2: (2) Random forest depends heavily on the training dataset. The authors do not provide an overview of the comprehensiveness of the training dataset: for instance, does the dataset cover all major chemical regimes in the EKMA plot, i.e., NOx-limited, VOC-limited, NO titration? The authors claim that ozone production in Beijing, China is mostly VOC-limited, which is consistent with previous studies. If the training set collected in Beijing does not have sufficient coverage in the NOx-limited regime, then the trained algorithm essentially attempts to extrapolate in that regime, which is dangerous and prone to overtraining. I would then question the if this random forest

model can make meaningful forecast in that regime at all.

Reply: Thank you for your valuable suggestion. We added the description of training dataset in Text S3 in the revised SI. This point has been replied in the aforementioned question. The mean values (±standard deviation) of the input and output parameters for the training data set are shown in Table R1. This Table was also added as Table S1 in the revised SI.

Table R1. An overview of training dataset from 2014 to 2016 during the observation period.

species / unit	2014				2015				2016			
	Measured VOC		Initial VOC		Measured VOC		Initial VOC		Measured VOC		Initial VOC	
	average	std. dev.*	average	std. dev.	average	std. dev.	average	std. dev.	average	std. dev.	average	std. dev.
Cyclopentane / ppbv	0.95	1.05	0.95	1.05	0.00	0.00	0.00	0.00	0.27	0.29	0.27	0.29
Ethane / ppbv	2.38	0.98	2.39	0.98	1.84	0.88	1.85	0.89	1.07	0.51	1.07	0.51
Acetylene / ppbv	1.64	1.31	1.65	1.31	0.13	0.33	0.14	0.33	0.32	0.30	0.32	0.30
Propane / ppbv	2.44	1.60	2.46	1.61	2.42	1.75	2.45	1.76	1.35	0.93	1.36	0.93
Benzene / ppbv	0.60	0.44	0.61	0.44	0.47	0.35	0.47	0.36	4.59	4.23	4.64	4.29
iso-Butane / ppbv	0.95	0.66	0.96	0.67	0.35	0.53	0.35	0.54	0.24	0.18	0.24	0.19
2,2-Dimethylbutane / ppbv	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00
n-Butane / ppbv	1.57	1.11	1.60	1.11	0.67	0.87	0.69	0.89	0.85	0.73	0.87	0.74
2,2,4-Trimethylpentane / ppbv	0.01	0.04	0.01	0.04	0.04	0.07	0.05	0.07	0.02	0.02	0.02	0.02
iso-Pentane / ppbv	0.11	0.38	0.11	0.40	0.00	0.00	0.00	0.00	0.16	0.18	0.16	0.18
2,3-Dimethylpentane / ppbv	0.07	0.08	0.07	0.08	0.06	0.08	0.06	0.08	0.02	0.03	0.02	0.03
3-Methylhexane / ppbv	0.06	0.07	0.06	0.07	0.04	0.05	0.04	0.05	0.01	0.02	0.01	0.02
Toluene / ppbv	1.28	1.02	1.32	1.04	0.88	1.55	0.93	1.57	0.30	0.34	0.32	0.37
2,3-Dimethylbutane / ppbv	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.06	0.08	0.06	0.08
n-Propyl benzene / ppbv	0.01	0.02	0.01	0.02	0.01	0.03	0.01	0.03	0.04	0.11	0.05	0.11

iso-Propyl benzene / ppbv	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.01	0.06
2,3,4-trimethylpentane / ppbv	0.12	0.29	0.12	0.31	0.06	0.10	0.06	0.11	0.01	0.02	0.02	0.02
n-hexane / ppbv	0.37	0.30	0.39	0.31	0.05	0.18	0.06	0.20	0.18	0.27	0.19	0.30
n-heptane / ppbv	0.08	0.09	0.09	0.10	0.06	0.06	0.06	0.07	0.02	0.02	0.02	0.02
2-methylhexane / ppbv	0.03	0.03	0.03	0.04	0.02	0.04	0.02	0.04	0.01	0.01	0.01	0.01
3-methylhexane / ppbv	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.02	0.00	0.01	0.00	0.01
cyclohexane / ppbv	0.04	0.05	0.05	0.05	0.03	0.05	0.04	0.05	0.04	0.10	0.04	0.12
ethylbenzene / ppbv	0.33	0.31	0.34	0.32	0.21	0.23	0.23	0.25	0.10	0.15	0.10	0.16
n-octane / ppbv	0.04	0.11	0.04	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ethene / ppbv	2.15	1.36	2.31	1.43	1.72	1.16	1.90	1.25	0.39	0.30	0.41	0.31
methylcyclohexane / ppbv	0.01	0.03	0.01	0.03	0.01	0.03	0.01	0.04	0.02	0.03	0.02	0.04
n-nonane / ppbv	0.03	0.04	0.03	0.04	0.02	0.02	0.02	0.03	0.02	0.04	0.02	0.04
n-decane / ppbv	0.02	0.04	0.02	0.05	0.02	0.03	0.02	0.03	0.00	0.01	0.00	0.01
p-ethyltoluene / ppbv	0.06	0.08	0.06	0.08	0.02	0.03	0.03	0.04	0.07	0.10	0.07	0.11
p-diethyl benzene / ppbv	0.01	0.04	0.01	0.04	0.01	0.02	0.01	0.02	0.09	0.17	0.11	0.22
o-ethyl toluene / ppbv	0.03	0.04	0.04	0.04	0.01	0.03	0.01	0.03	0.08	0.28	0.09	0.32
o-xylene / ppbv	0.09	0.18	0.10	0.18	0.16	0.18	0.19	0.20	0.14	0.26	0.15	0.27
m-ethyl toluene / ppbv	0.02	0.07	0.02	0.07	0.04	0.09	0.04	0.09	0.03	0.04	0.03	0.05
m-diethyl benzene / ppbv	0.01	0.03	0.01	0.03	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.02
m/p-Xylene / ppbv	0.61	0.64	0.68	0.65	0.45	0.51	0.54	0.59	0.22	0.38	0.25	0.41
propene / ppbv	2.07	1.18	2.83	2.26	4.40	2.61	6.60	6.12	0.28	0.41	0.34	0.45
1-Butene / ppbv	0.10	0.14	0.13	0.17	0.04	0.10	0.08	0.25	0.03	0.03	0.04	0.06
1-Pentene / ppbv	0.03	0.09	0.04	0.09	0.03	0.07	0.05	0.12	0.02	0.06	0.02	0.07
1,2,4-trimethyl benzene/ ppbv	0.01	0.08	0.01	0.08	0.08	0.09	0.11	0.12	0.05	0.05	0.06	0.09
1,2,3-trimethyl benzene/ ppbv	0.00	0.01	0.00	0.01	0.03	0.05	0.04	0.08	0.05	0.28	0.05	0.28
a-pinene / ppbv	0.01	0.03	0.02	0.03	0.01	0.03	0.01	0.03	0.18	0.46	0.84	3.48
cis-2-Butene / ppbv	0.34	0.70	0.85	2.67	0.66	0.85	1.77	4.56	0.04	0.05	0.11	0.29

1,3,5-Trimethylbenzene / ppbv	0.05	0.07	0.08	0.11	0.03	0.05	0.07	0.14	0.25	0.56	1.07	4.11
styrene / ppbv	0.18	0.27	0.30	0.61	0.00	0.03	0.01	0.08	0.27	0.79	0.57	2.08
2-methyl-1-pentene / ppbv	0.18	0.37	0.72	2.94	0.04	0.04	0.26	1.68	0.02	0.09	0.03	0.12
trans-2-Butene / ppbv	0.08	0.16	0.24	1.15	0.09	0.11	0.34	0.74	0.02	0.02	0.04	0.08
cis-2-Pentene / ppbv	0.15	0.20	0.37	0.93	0.17	0.17	0.91	4.24	0.01	0.02	0.02	0.08
1,3-Butadiene / ppbv	0.09	0.10	0.19	0.34	0.04	0.05	0.12	0.38	0.02	0.03	0.05	0.25
trans-2-Pentene / ppbv	0.03	0.08	0.06	0.27	0.01	0.02	0.11	0.89	0.01	0.02	0.01	0.05
β-pinene / ppbv	0.00	0.01	0.01	0.03	0.01	0.01	0.02	0.15	0.00	0.01	0.00	0.02
isoprene / ppbv	0.89	0.64	5.70	18.78	0.34	0.43	6.40	21.56	0.13	0.17	2.12	7.46
NO / ppbv	7.03	17.02	7.03	17.02	3.38	5.59	3.38	5.59	5.28	10.35	5.28	10.35
NO_2 / ppbv	15.50	15.79	15.50	15.79	19.11	12.68	19.11	12.68	18.72	12.40	18.72	12.40
T / ° C	22.56	6.28	22.56	6.28	22.70	5.24	22.70	5.24	22.37	4.85	22.37	4.85
RH / %	50.93	23.88	50.93	23.88	41.49	23.23	41.49	23.23	36.23	21.58	36.23	21.58
SR / W m^{-2}	162.92	222.95	162.92	222.95	153.29	205.01	153.29	205.01	150.81	199.35	150.81	199.35
WS / m s^{-1}	3.11	2.70	3.11	2.70	2.29	2.15	2.29	2.15	1.25	1.24	1.25	1.24
WD / °	162.42	105.07	162.42	105.07	175.38	101.87	175.38	101.87	184.21	108.06	184.21	108.06
$PM_{2.5}$ /μg m^{-3}	67.16	53.47	67.16	53.47	63.13	56.46	63.13	56.46	61.05	48.64	61.05	48.64
CO /mg m^{-3}	0.78	0.49	0.78	0.49	0.68	0.44	0.68	0.44	0.57	0.36	0.57	0.36
O_3 / ppbv	44.32	32.38	44.32	32.38	42.74	27.94	42.74	27.94	44.01	29.64	44.01	29.64

* Standard Deviation (std. Dev.)

In Text S3 in the revised SI, we added a short paragraph "Data description. The length of the input data from 2014 to 2016 were 1190, 1062 and 872 rows, respectively, in which different types of VOCs, NOx, CO, $PM_{2.5}$ and meteorological parameters (including temperature, relative humidity, solar radiation, wind speed and direction)

were used as input variables and O_3 was the output variable. The mean values (±standard deviation) of input/output parameters are shown in Table S1"

As shown in Figure R1 or Figure S15, the training dataset were located in VOC-limited, NOx-limited, and transition regimes, while most of the training data were located in the VOC-limited regime. To avoid overtraining, we performed a 12-fold cross-validation, i.e., by randomly dividing the observed data into 12 subsets and alternately taking one subset as testing data and the rest as training data, to ensure that each data point has an equal chance of being trained and tested. Figure R2 (Figure 2 in the revised manuscript) shows the comparisons between the measured and predicted O_3 concentrations using different VOC inputs. The curves of the predicted O_3 concentrations were spliced using the testing datasets in all runs. Thus, both the training data and the testing data actually covered all the sensitivity regimes of O_3 formation. We think that the model is robust in the revised version according to your good suggestion.

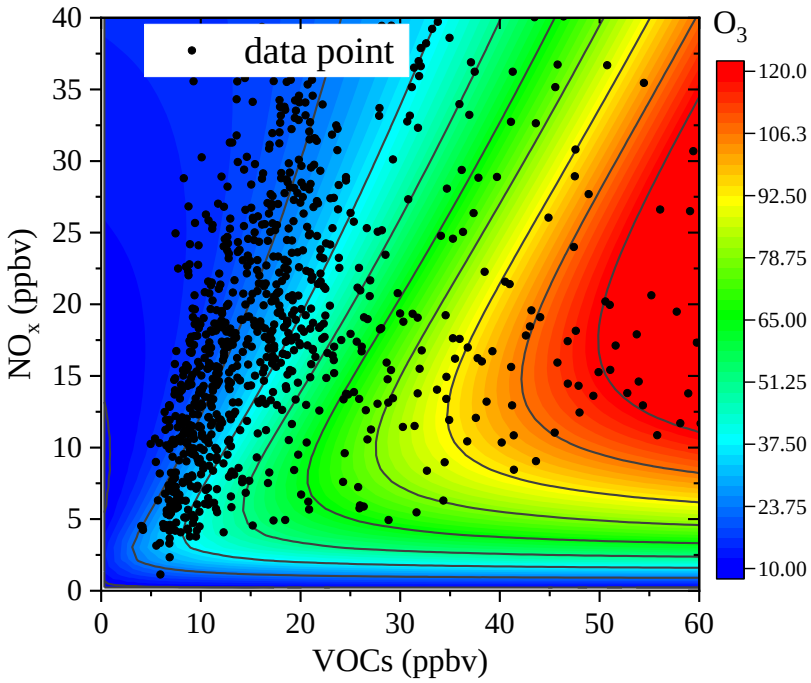

Figure R1. Sensitivity curves of O_3 formation and distribution of training data in 2015.

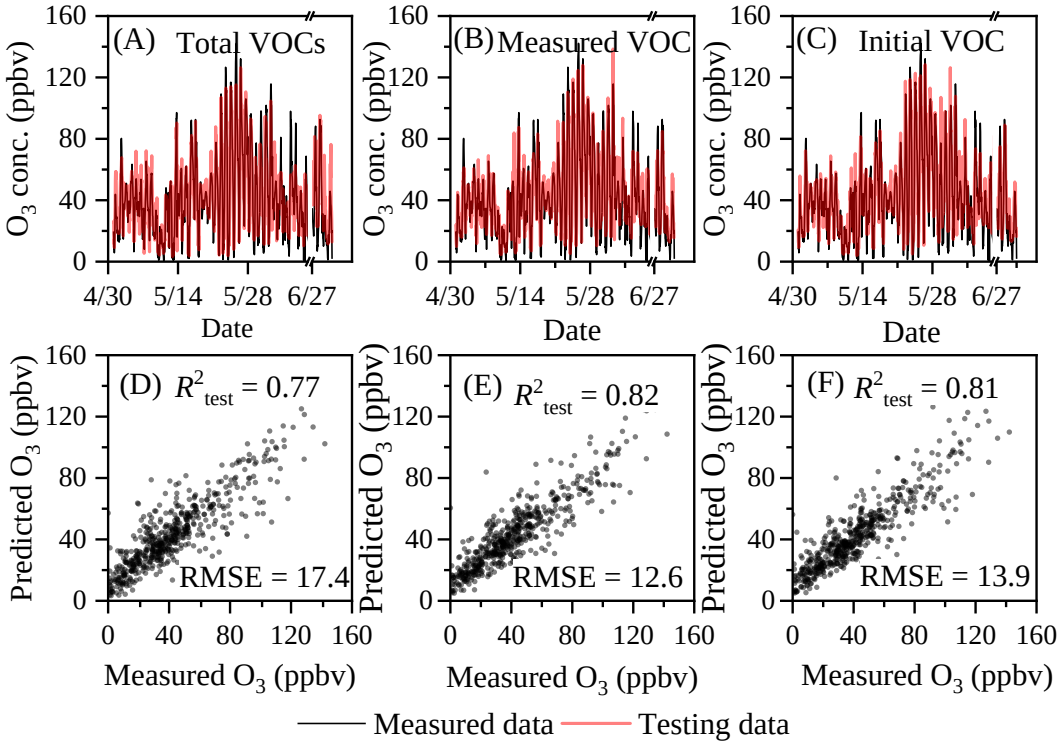

Figure R2. Comparison of the predicted and measured O_3 concentrations in Beijing in the summer of 2015. (A and D: TVOC concentrations; B and E: measured concentrations of VOC species; C and F: initial concentrations of VOC species). The testing data curves were spliced using the testing datasets in all runs during the 12-fold cross-validation.

When plotting the O_3 formation sensitivity curves, we made a virtual matrix of inputs by varying the concentrations of NOx and VOCs from 0.9 to 1.1 times (with a step of 0.01) of their mean values (observed point data) while keeping all other inputs unchanged (i.e., the mean values during our observations). Then, the new matrix was used as testing data, while all the measured data were taken as training data. Thus, the testing data should represent the mean sensitivity regime of O_3 in Beijing, while the training data actually covered all the sensitivity regimes of O_3 formation. As shown in Figure R3 or Figure 4B, the sensitivity of O_3 formation was located in VOC-limited regime in 2015. The mean relative errors of simulated O_3 between RF model and Box model was 15.6% (Figure R4 or Figure S8), which means that the RF model can well predict the O_3 concentrations and the sensitivity regime of O_3 formation. Although we

set relatively small variations of VOCs and NOx (±10%) compared to the observed values to decrease the model uncertainty when depicting the EKMA curves, the training data represent the real conditions in Beijing during our observations. Therefore, we think our results should be reliable and meaningful.

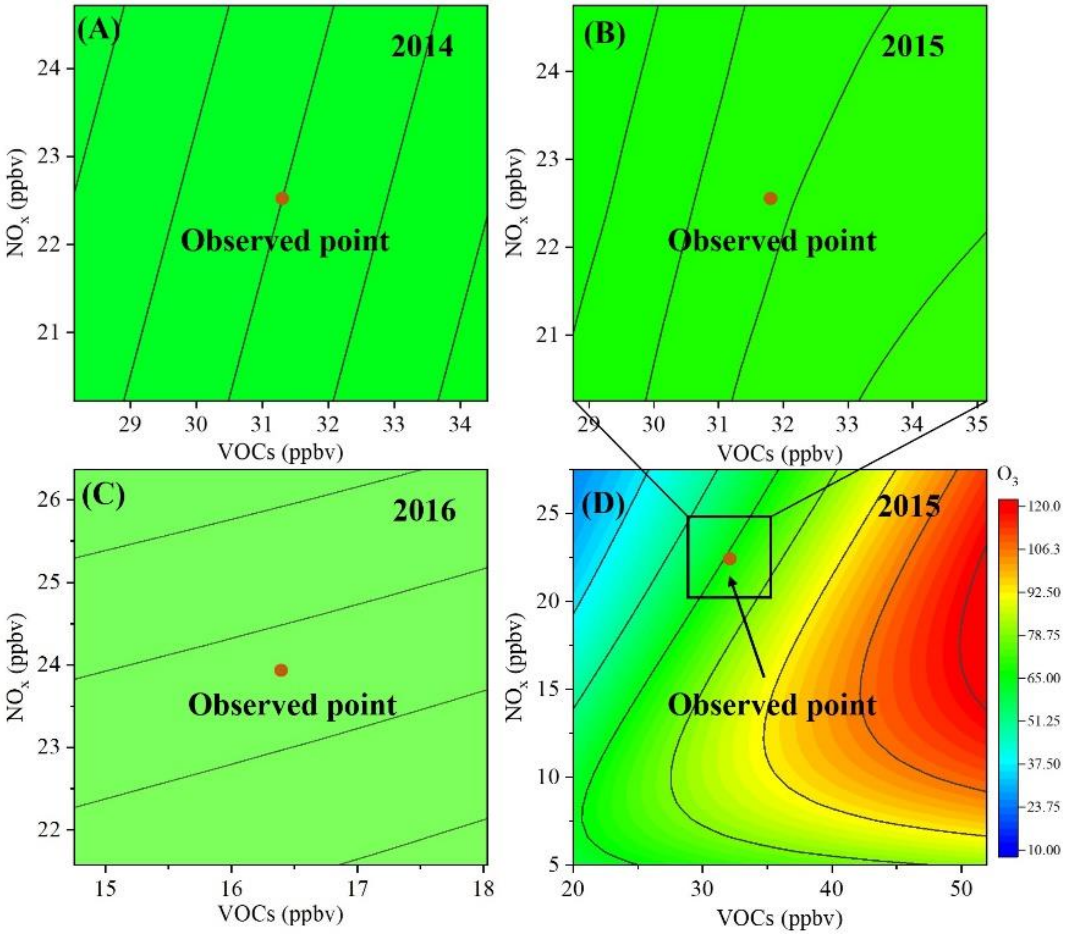

Figure R3. Ozone formation sensitivity curves from 2014-2016. (A, B, C: calculated by the RF model for 2014, 2015, and 2016, respectively. D: calculated by the OBM for 2015)

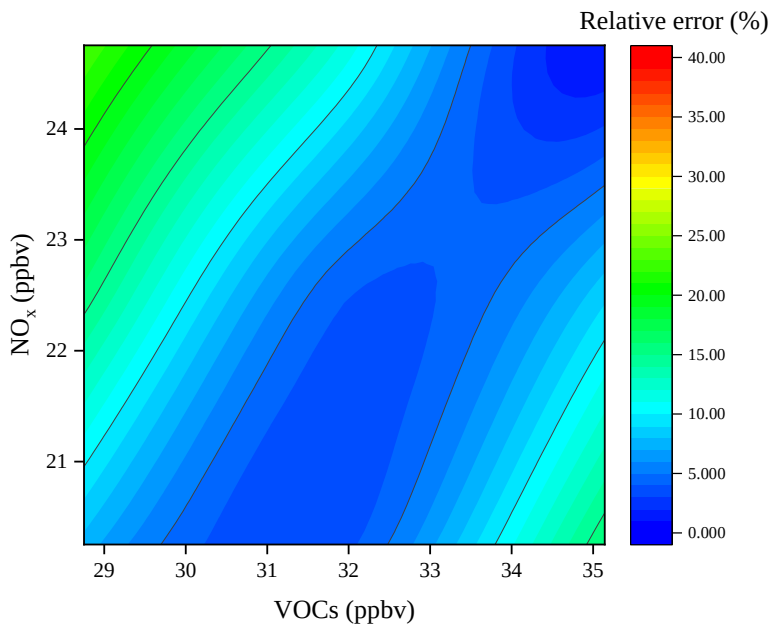

Figure R4. The relative error of simulated O_3 concentrations between the RF model and the box model in 2015.

In lines 232-240 the revised manuscript, we added a paragraph "It should be pointed out that if the training dataset does not have sufficient coverage in the NOx-limited regime, then the trained algorithm essentially attempts to extrapolate in that regime, which is prone to overtraining. To avoid such overtraining, a 12-fold cross-validation by randomly dividing the observation data in each day into 12 subsets and alternately taking one subset as testing data and the rest as training data ensures that each data point has an equal chance of being trained and tested. The curves of the predicted O_3 concentrations in Figure 2 were spliced using the testing datasets in all runs. Thus, our results actually covered all the sensitivity regimes of O_3 formation. This means that the model is robust".

In lines 171-178 in the revised manuscript, we added a paragraph "When plotting the O_3 formation sensitivity curves, we made a virtual matrix of inputs by varying the concentrations of NOx and VOCs from 0.9 to 1.1 times (with a step of 0.01) of their mean values while keeping all other inputs unchanged (i.e., the mean values). Then, the new matrix was used as testing data, while all the measured data were taken as training data. Thus, the testing data should represent the mean sensitivity regime of O_3 in Beijing, while the training data actually covered all the sensitivity regimes of O_3 formation to

guarantee a sufficient coverage in the NOx-limited regime for the RF model simulations".

Q3: (3) The calculation of the initial VOC concentrations is problematic: the method depends heavily on initial/source ratio which is not discussed at all in this work; the method assumes biogenic VOCs share the same air mass histories as the anthropogenic VOCs which is not supported by any evidence. For these reasons, I do not recommend the current form of the manuscript for publication in Atmospheric Measurement and Techniques. Given the substantial amount of work needed to demonstrate the robustness of the machine learning workflow, to outline key details in a transparent manner, and to revise the initial VOC calculation, resubmission is recommended. Please see my specific and minor/technical comments below.

Reply: Thank you so much for your good suggestions. In previous studies (Shao et al., 2011; Zhan et al., 2021), it has been justified for selecting the pair of ethylbenzene/xylene as the tracers when calculating ambient OH exposure in terms of the following rules: 1) the concentrations of xylene and ethylbenzene are well correlated, which indicates that they are simultaneously emitted; 2) they have different degradation rates in the atmosphere; 3) the calculated initial VOCs are in good agreement with those calculated using other tracers, such as toluene/benzene.

As shown in Figure R5 or Figure S9, the concentrations of xylene and ethylbenzene correlated well during our observations in this work. In addition, we compared the ratio of the initial concentrations calculated according to the ratio of xylene/ethylbenzene with that using the ratio of toluene/benzene (Figure R6 or S9). Except for several compounds, the ratio of the PICs for most of these VOCs varied within 1.0±0.1. This means the calculated photochemical initial concentrations (PICs) are in good agreement when using different tracers. Sensitivity tests showed that the uncertainty caused by the OH exposure (from −10% to +10%) ranged from 0.55 to 1.57 (Table R2 or Table S4). Figure R7 or Figure S12 shows the calculated diurnal curves of the PICs from 2014 to 2016. Photochemical losses of VOCs occurred prominently during the daytime.

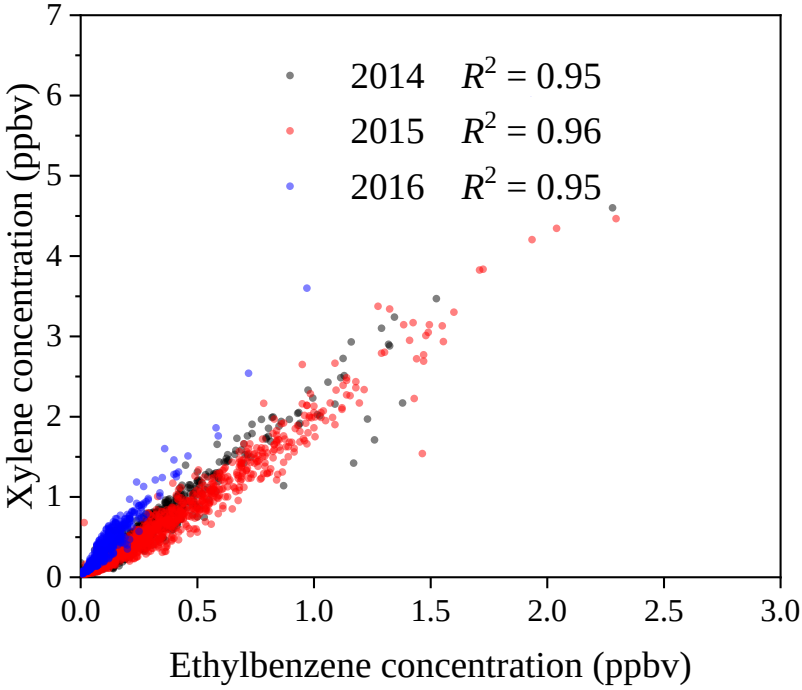

Figure R5. The relationship between xylene and ethylbenzene.

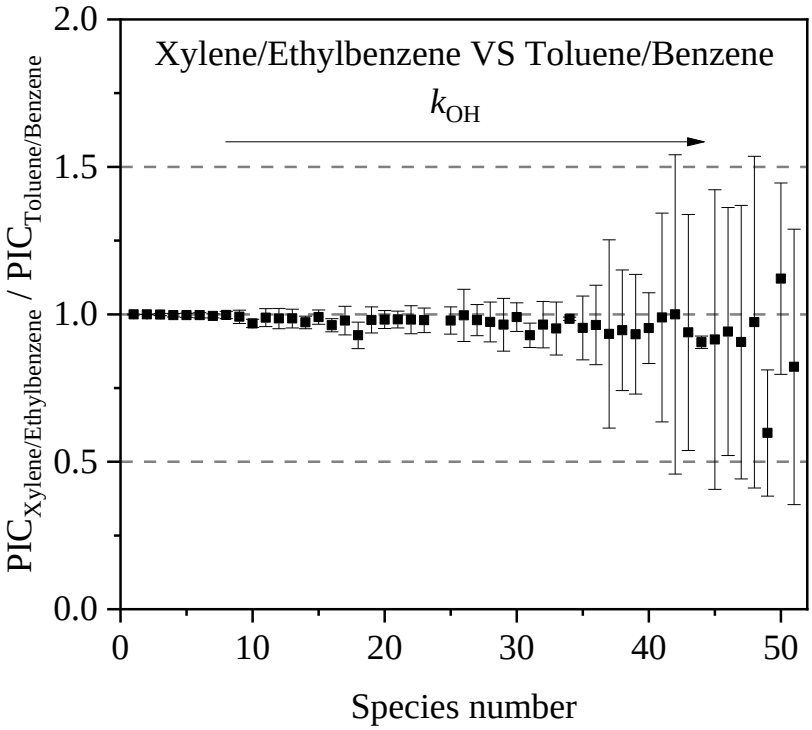

Figure R6. Comparison of the initial VOCs calculated using the ratio of xylene/ethylbenzene with that using the ratio of toluene/benzene in 2015. (Error bars are standard deviations.)

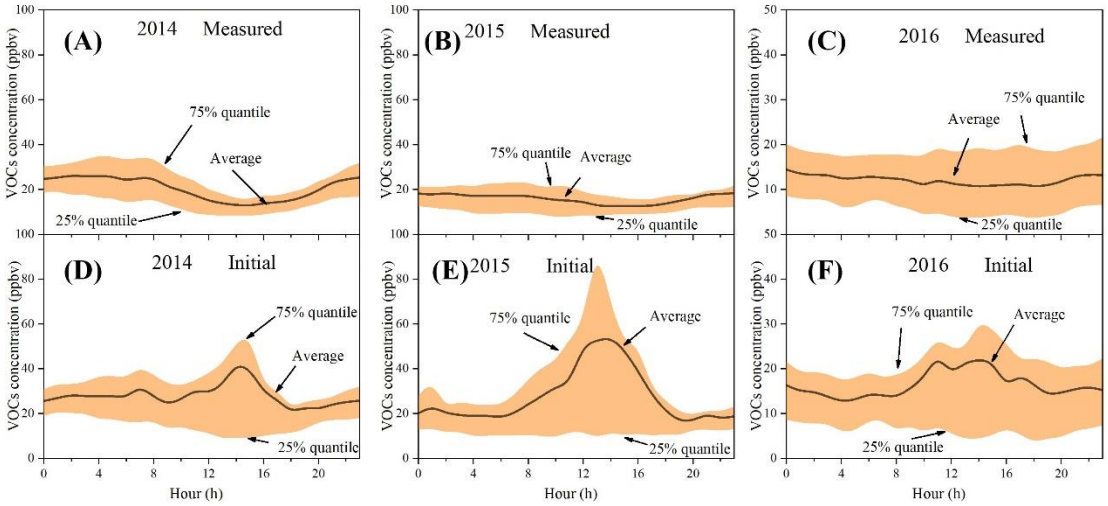

Figure R7. The daily variation of VOCs concentration. (A and D for 2014; B and E for 2015; C and F for 2016)

Table R2. k_{OH}, Method Detection Limit (MDL) and sensitivity test on estimation of [OH]×t of different VOC species

Specie number	species name	k_{OH}*	MDL**	Ratio to the initial VOC***					
				2014		2015		2016	
				-10% [OH]×t	+10% [OH]×t	-10% [OH]×t	+10% [OH]×t	-10% [OH]×t	+10% [OH]×t
1	Ethane	0.254	0.050	1.00	1.00	1.00	1.00	1.00	1.00
2	Acetylene	0.756	0.022	1.00	1.00	1.00	1.00	1.00	1.00
3	Propane	1.11	0.013	1.00	1.00	1.00	1.00	1.00	1.00
4	Benzene	1.22	0.011	1.00	1.00	1.00	1.00	1.00	1.00
5	iso-Butane	2.14	0.010	1.00	1.00	1.00	1.00	1.00	1.00
6	2,2-Dimethylbutane	2.27	0.005	1.00	1.00	1.00	1.00	1.00	1.00
7	n-Butane	2.38	0.011	1.00	1.00	1.00	1.00	1.00	1.00
8	2,2,4-Trimethylpentane	3.38	0.008	1.00	1.00	1.00	1.00	1.00	1.00
9	iso-Pentane	3.6	0.008	1.00	1.00	1.00	1.00	1.00	1.00

10	Cyclopentane	5.02	0.005	1.00	1.00	1.00	1.00	1.00	1.00
11	n-hexane	5.25	0.011	0.99	1.01	0.99	1.01	0.99	1.01
12	Toluene	5.58	0.009	1.00	1.00	0.99	1.01	1.00	1.00
13	2,3-Dimethylbutane	5.79	0.004	1.00	1.00	1.00	1.00	0.99	1.01
14	n-Propyl benzene	5.8	0.008	1.00	1.00	1.00	1.00	0.99	1.01
15	iso-Propyl benzene	6.3	0.007	1.01	1.01	0.99	1.01	0.97	1.03
16	2,3,4-trimethylpentane	6.6	0.008	0.99	1.01	0.99	1.01	1.00	1.00
17	n-heptane	6.81	0.009	0.99	1.01	0.99	1.01	0.99	1.01
18	ethylbenzene	7	0.009	0.99	1.01	0.99	1.01	0.99	1.01
19	cyclohexane	7.02	0.011	1.00	1.00	0.99	1.01	0.99	1.01
20	2,3-Dimethylpentane	7.15	0.009	1.00	1.00	1.00	1.00	1.00	1.00
21	3-Methylhexane	7.17	0.009	1.00	1.00	0.99	1.01	1.00	1.00
22	ethene	8.15	0.021	0.99	1.01	0.99	1.01	0.99	1.01
23	n-octane	8.16	0.008	0.99	1.01	1.00	1.00	1.00	1.00
24	2-Methylheptane	8.31	0.008	1.00	1.00	0.99	1.01	0.99	1.01
25	3-Methylheptane	8.59	0.008	1.00	1.00	1.00	1.01	0.99	1.01
26	methylcyclohexane	9.64	0.005	0.99	1.01	0.99	1.01	0.99	1.01
27	n-nonane	9.75	0.006	0.99	1.01	0.99	1.01	0.98	1.02
28	n-decane	11	0.007	0.99	1.01	0.99	1.01	0.99	1.01
29	p-ethyl toluene	11.8	0.007	0.99	1.01	0.98	1.02	0.98	1.02
30	p-diethyl benzene	-	0.008	1.00	1.00	0.99	1.01	0.97	1.03
31	o-ethyl toluene	11.9	0.007	0.99	1.01	0.99	1.01	1.00	1.00
32	o-xylene	13.6	0.007	0.99	1.01	0.98	1.02	1.00	1.00
33	m-ethyl toluene	18.6	0.010	0.99	1.01	0.99	1.01	0.97	1.03
34	m-diethyl benzene	-	0.009	0.99	1.01	0.99	1.01	0.98	1.02
35	m/p-Xylene	23.1/14	0.008	0.99	1.01	0.98	1.02	0.98	1.03

		.2							
36	propene	26	0.015	0.96	1.04	0.95	1.05	0.96	1.05
37	1-Butene	31.1	0.010	0.97	1.04	0.90	1.12	0.92	1.10
38	1-Pentene	31.4	0.009	0.98	1.02	0.93	1.09	0.93	1.08
39	1,2,4-trimethyl benzene	32.5	0.008	1.00	1.01	0.95	1.05	0.91	1.10
40	1,2,3-trimethyl benzene	32.7	0.009	0.96	1.04	0.96	1.04	0.97	1.03
41	a-pinene	51.8	0.010	0.97	1.04	0.96	1.05	0.75	1.35
42	cis-2-Butene	55.8	0.019	0.87	1.16	0.86	1.17	0.77	1.32
43	1,3,5-Trimethylbenzene	56.7	0.007	0.93	1.08	0.90	1.13	0.73	1.37
44	styrene	58	0.010	0.91	1.11	0.90	1.13	0.98	1.02
45	2-methyl-1-pentene	63	0.002	0.81	1.25	0.70	1.49	0.81	1.28
46	trans-2-Butene	63.2	0.014	0.84	1.22	0.82	1.25	0.76	1.35
47	cis-2-Pentene	65	0.006	0.86	1.19	0.74	1.42	0.83	1.24
48	1,3-Butadiene	65.9	0.024	0.88	1.16	0.82	1.26	0.87	1.18
49	trans-2-Pentene	67	0.009	0.88	1.16	0.63	1.63	0.75	1.38
50	β-pinene	73.5	0.010	0.90	1.12	0.81	1.26	0.92	1.10
51	isoprene	99.6	0.009	0.73	1.40	0.67	1.50	**0.55**	**1.57**

* Unit: 10^{-12} cm^3 $mole^{-1}$ s^{-1}. k_{OH} values were under conditions of 300K. (Carter 2010)

** Unit: ppb. The relative standard derivations (RSDs) were within 10% for the target compounds in all six replicates.

*** All species were selected for sensitivity tests of initial VOCs to [OH]×t. The reaction rates of these species with OH covered the range of 51 VOCs and were characterized by low, medium and high k_{OH} levels. The sensitivity test results showed that the uncertainty in the estimation of initial VOCs caused by the [OH]×t estimation uncertainty ranged from 0.55 to 1.57.

Potential source contribution function (PSCF) analysis has been further carried out to evaluated the possible influence of air mass on the emission ratio of ethylbenzene

and xylene. As shown in Figure R8A-D or Figure S11, xylene showed similar pattern to ethylbenzene in the early morning or in the whole day. These results indicate that variations of air mass should have little influence on their initial ratio. In addition, isoprene showed similar patterns to that of xylene and ethylbenzene (Figure R8G-H), which means VOC emissions are evenly distributed in Beijing. This can be ascribed to the fact that our observation site is a typical urban station. Although isoprene and xylene/ethylbenzene are from biogenic sources and anthropogenic sources, both them are non-point sources on a city scale.

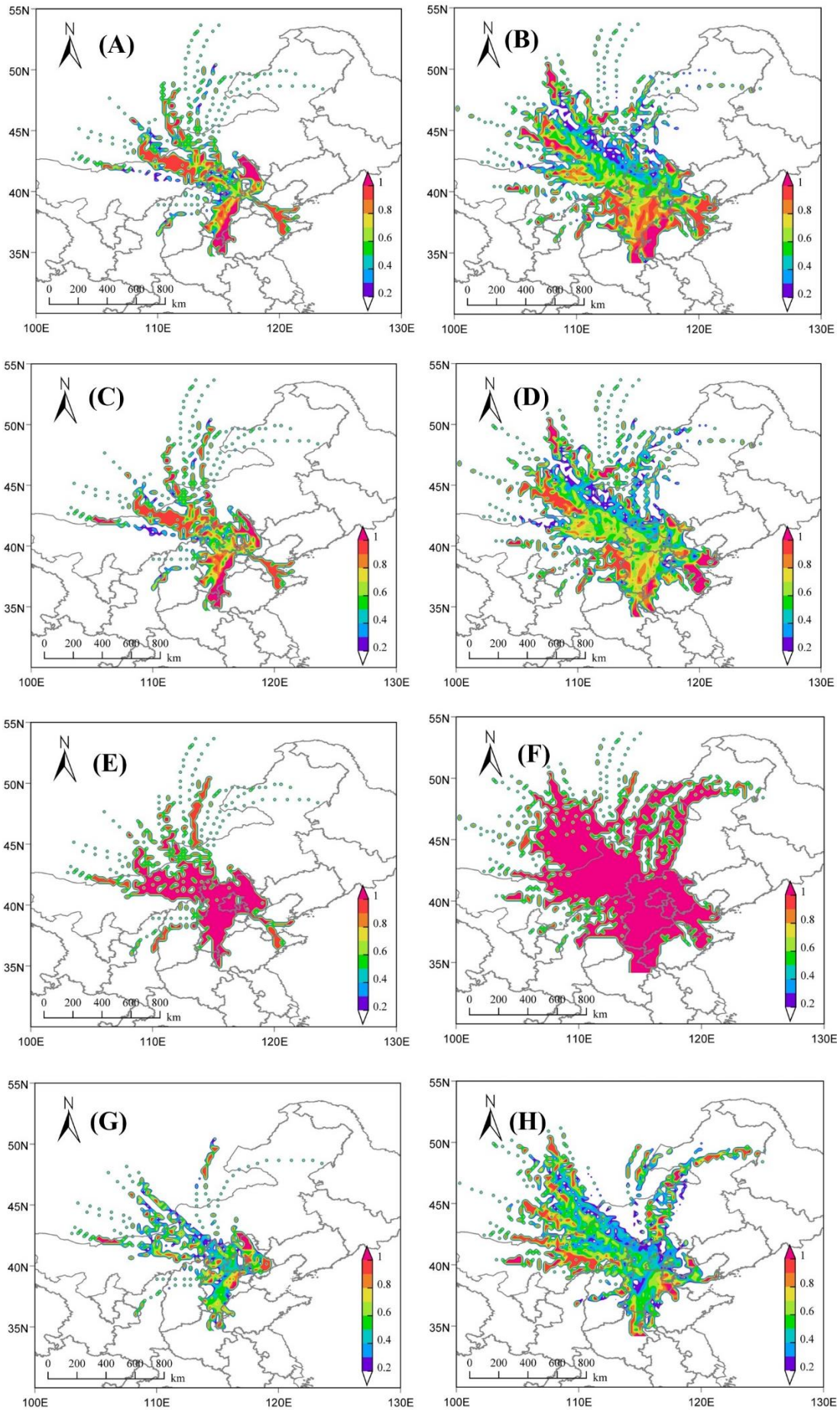
(A)
(B)
(C)
(D)
(E)
(F)
(G)
(H)
N
55N
50N
45N
40N
35N
100E
110E
120E
130E
0 200 400 600 800
km
1
0.8
0.6
0.4
0.2

Figure R8. The potential source contribution function (PSCF) maps for ethylbenzene (A and B), xylene (C and D), ratio of xylene to ethylbenzene (E and F), and isoprene (G and H) arriving in the observation site. The figures A, C, E and G are the results for the morning (05:00 and 06:00), and the figures of B, D, F and H are the results of the whole day (00:00-23:00).

In the revised SI, we added the details on calculation of the initial VOCs and more discussions in the Text S2. The changes are shown below.

"**Text S2. Calculation of initial VOCs concentrations**

Photochemical initial concentration (PIC) proposed by Shao et al. (2011), which is calculated based on the photochemical-age approach and has been applied to evaluate the effect of photochemical processing on measured VOC levels. Equation S1 essentially describes the integrated OH exposure (Shao et al., 2011).

$$\int c_{OH}dt = \frac{1}{k_{A,OH}-k_{B,OH}}\left[\ln\left(\frac{VOC_A}{VOC_B}\right)_{initial} - \ln\left(\frac{VOC_A}{VOC_B}\right)\right] \quad (S1)$$

The initial concentration of species *i* can be calculated using Equation S2.

$$VOC_{i,\,initial} = \frac{VOC_i}{\exp(-k_{i,OH})\exp(\int c_{OH}dt)} \quad (S2)$$

Substituting equation 1 into equation 2, then we can get equation S3.

$$VOC_{i,\,initial} = \frac{VOC_i}{\exp(-k_{i,\,OH})\exp\left(\frac{1}{k_{A,OH}-k_{B,\,OH}}\left[\ln\left(\frac{VOC_A}{VOC_B}\right)_{initial} - \ln\left(\frac{VOC_A}{VOC_B}\right)\right]\right)} \quad (S3)$$

Where C_{OH} represents the ambient OH concentration; $k_{A,OH}$ and $k_{B,OH}$ represent the reaction rate of compound A and B with OH radical, respectively; t represents the reaction time of species *i* in the ambient.

In previous work (Shao et al., 2011; Zhan et al., 2021), the selection of ethylbenzene and xylene as tracers was justified for calculating ambient OH exposure under the following conditions: 1) the concentrations of xylene and ethylbenzene were well correlated (Figure S9), which indicated that they were simultaneously emitted; 2) they had different degradation rates in the atmosphere; and 3) the calculated PICs were in good agreement with those calculated using other tracers (Shao et al., 2011; Zhan et al., 2021).

In this study, the ethylbenzene/xylene pair was used to calculate ambient OH

exposure. As shown in Figure S9, the concentrations of xylene and ethylbenzene are well correlated, which indicates that they are simultaneously emitted. In addition, we compared the PICs according to xylene/ethylbenzene with that using toluene/benzene (Figure S10). The calculated PICs ratio ($PIC_{Xylene/Ethylbenzene}$ / $PIC_{Toluene/Benzene}$) varied from 0.5 to 1.5 with a mean value of 0.96. This means the calculated initial VOCs was in good agreement when using different tracers. The mean ratio (0.52, from 0.45 to 0.66) of ethylbenzene/xylene before sunrise was taken as the initial ratio of ethylbenzene/xylene. Sensitivity tests showed that the uncertainty of PICs caused by the OH exposure (from −10% to +10%) ranged from 0.55 to 1.57 (Table S4).

Variations of air mass may also affect the VOC ratio. Figure S11 A-D shows the mean concentration distribution of ethylbenzene and xylene in the early morning and the whole day based on potential source contribution function (PSCF) analysis. Xylene showed similar patterns to ethylbenzene in different air mass trajectories and different periods. These results indicate that the emissions of xylene and ethylbenzene were constant throughout the day and variations of air mass should have little influence on the initial ratio of VOCs. The hourly concentrations of ethylbenzene and xylene were used to calculate the concentration of initial VOCs. The initial VOC was calculated by adding the measured VOC concentration and the calculated photochemical loss. Figure S12 shows the diurnal variations of the observed and initial VOCs concentrations from 2014 to 2016. Photochemical loss of VOC occurred mainly during the daytime.

It should be noted that the lifetimes ($1/k_2c_{OH}$) of highly reactive VOCs, such as isoprene, greatly depend on the OH exposure. The photochemical ages of isoprene were 0.01–6.21 h (1.26 ± 1.12 h). This value is comparable with previously reported photochemical ages (Shao et al., 2011; Gao et al., 2018). However, the initial concentrations of highly reactive VOCs may be overestimated due to their short lifetimes and should be taken as the upper limits. On the other hand, isoprene is a biogenic VOC, while xylene and ethylbenzene are anthropogenic VOCs. If they do not share the same air mass histories, an additional uncertainty is inevitable for the PICs of isoprene. As shown in Figure S11, isoprene showed similar patterns to that of xylene and ethylbenzene, which means VOC emissions are evenly distributed in Beijing during

our observations. This can be ascribed to the fact that our observation site is a typical urban station. Although isoprene and xylene/ethylbenzene different sources, both them are non-point sources on a city scale. Therefore, the photochemical clock calculated using xylene and ethylbenzene is able to correct the photochemical loss of biogenic VOCs to some extent. It should be noted that uncertainty is inevitable when we estimating the photochemical age (Parrish et al., 2007). However, the aim of this work is to test whether the ML-model can reflect the influence of photochemical loss of VOCs species on O_3 modelling. The PICs should provide additional information for understanding O_3 formation in the atmosphere.".

As for the robustness and workflow of RF model, we have replied this point in **Q1** and **Q2** and revised the manuscript and the SI accordingly.

Specific concerns:

Q4: Line 70-71: This is a valid concern. However, the machine learning based approaches are also subject to this.

Reply: Thank you for your comment. We have added a paragraph to point out the shortcoming of machine learn "Although attentions should be paid to the robustness of machine learning because it depends on input dataset (observations or outputs from chemical transport models), previous studies have demonstrated that cross-validation and data-normalization can well reduce the dependence of the model on input data and improve the robustness of the model (Wang et al., 2016; Wang et al., 2017b; Liu et al., 2021; Ma et al., 2021)" in Lines 79-83 in the revised manuscript.

Q5: Line 72-73: Respectfully, I disagree. Box models using condensed mechanisms are usually quite cheap. Near-explicit mechanisms such as MCM are more expensive, but the EKMA-type configurations are still considerably cheaper than 3D chemical transport models. Well-developed box models with MCM or other condensed mechanisms configured for ozone sensitivity (EKMA) can run on personal computers or small servers/clusters, providing timely predictions with no major demand for computational resources. OBMs generally are not considered as being time-consuming

or computationally expensive.

Reply: Thank you. We have deleted the sentences "In addition, both of them are time-consuming and expensive when computational resources are considered" and "Traditional models have difficulty assessing O_3 formation sensitivity in a timely manner due to the limitations of flexibility and computational efficiency" in the revised manuscript.

Q6: Line 79-90: This section lists a few previous studies with vaguely portrayed methodologies and outcomes, but failed to mention any disadvantages of machine learning, such as the demand for large volume of comprehensive and good quality data, and of course the risk of overtraining. This section also fails to address a concern brought up earlier by the authors themselves: uncertainties and biases in the input dataset (observations, or outputs from chemical transport models). Please revise this section and discuss the applications of machine learning in air quality studies in the context of its disadvantages. Please also address how the impacts of input data (e.g., uncertainties and biases) might be reduced.

Reply: Thank you. We have revised this section and added the discussion in lines 79-86 in the reversed manuscript "Although attentions should be paid to the robustness of machine learning because it depends on input dataset (observations or outputs from chemical transport models), previous studies have demonstrated that cross-validation and data-normalization can well reduce the dependence of the model on input data and improve the robustness of the model (Wang et al., 2016; Wang et al., 2017b; Liu et al., 2021; Ma et al., 2021). Thus, it is a promising alternative to account for the effects of meteorology on air pollutants and has been intensively used in atmospheric study (Liu et al., 2020; Hou et al., 2022).".

Q7: Line 135-137: "… and then averages the scores of each decision tree as its final score…" This is a very vague description of the algorithm. Indeed, the ensemble prediction made by the entire forest is usually more accurate and robust than the predictions made by individual trees, relatively speaking. However, this (averaging

across all the decision trees) ABSOLUTELY DOES NOT guarantee that large biases and overfitting can be avoided. The splitting might help with mitigating the risk of overfitting but it is still FAR FROM BEING SUFFICIENT to guarantee the algorithm is not overfitted. Generally, much more comprehensive and rigorous measures than what is shown in this work are needed, for instance, multifold cross validation is a good idea. To further test the robustness of the machine learning workflow in real-world physics-driven problems, sometimes it is recommended to perform the cross validation with each fold being the data from a specific time period or geographic region.

Reply: Thank you so much for your good suggestion. We agree with you that the splitting might help with mitigating the risk of overfitting but it is still not sufficient to guarantee the algorithm is not overfitted. As replied in above questions, we performed multifold cross-validation according to your suggestion. A 12-fold cross-validation was carried out to mitigate the risk of overfitting. Briefly, we randomly divided the normalized data into 12 subsets, then alternately took one subset as testing data along with the rest as training data. By doing this, every data point has an equal chance being trained and tested.

In lines 145-153 in the revised manuscript, we have revised the sentence "… and then averages the scores of each decision tree as its final score…" to "The random forest (RF) is a type of ensemble decision tree that can be used for classification and regression (Breiman 2001). During the training process, the model creates a large number of different decision trees with different sample sets at each node, and then averages the results of all decision trees as its final results (Breiman 2001). To avoid over-fitting, we trained the random forest model using cross-validation for the normalized data, which can improve the robustness of the model. Briefly, we randomly divided the normalized data into 12 subsets, then alternately took one subset as testing data along with the rest as training data. By doing this, every data point has an equal chance being trained and tested.".

Q8: Line 132-: this section does not provide any details on whether the authors have performed any sort of hyperparameter tuning, which is important. How can the readers

be convinced that the performance is optimized? Information outlined in Table S1 is not at all sufficient to described how the algorithm is configured. And frankly, certain information in that table is practically useless (e.g., the method is "regression", and the sampling is "random"). Please also clarify if the authors implemented random forest by themselves or used that from a certain package (e.g., R, python). If latter, please specify which package is used.

Reply: Thank you for you good comments. In the revised SI, we added the process of hyperparameter tuning, such as the influence of the min sample split (Figure R9 or Figure S13) and tree number (Figure R10 or Figure S14) on the mean squared error (MSE). We also revised Table S2 (Table R3). We implemented random forest simulation by ourselves. The code was written by ourselves based on a MATLAB platform.

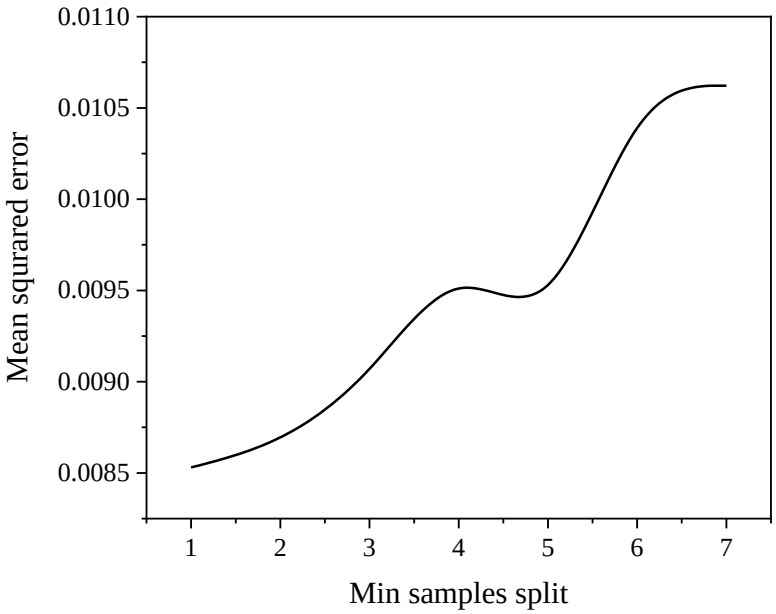

Figure R9. The relationship between min sample split and mean squared error (MSE)

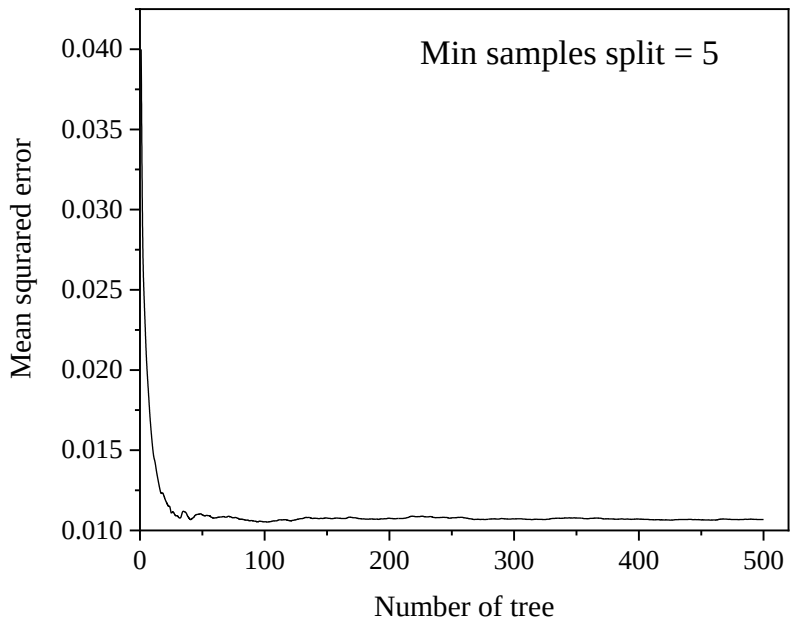

Figure R10. The relationship between trees number and mean squared error (MSE).

Table R3. RF model parameters and input parameters

RF model parameters		RF model inputs		
Type	Value	Type	input parameter	output parameter
Cross-validation	12	Figure 2A	Total VOC concentration, $PM_{2.5}$, NO, NO_2, CO, SR, RH, WD, WS, T	O_3
tree number	500	Figure 2B	51 measured VOC species, $PM_{2.5}$, NO, NO_2, CO, SR, RH, WD, WS, T	O_3
min sample split	5			
min sample leaf	1	Figure 2C	51 initial VOC species, $PM_{2.5}$, NO, NO_2, CO, SR, RH, WD, WS, T	O_3

Note: In this study, we optimized the number of tree and min samples split as shown in Figure S14 and S13, respectively. The min sample leaf was set 1 (default), and other parameter were set to default (auto/none).

Min samples split: the minimum number of samples to be split; Min samples leaf: the minimum number of samples in a leaf; Trees number: the number of trees during the training.

In lines 166-168 in the revised manuscript, we added a sentence "More details about workflow of RF model and the hyperparameter tuning can be found in the Text S3. The optimized parameters are shown in Table S2".

In the revised SI, we added a short paragraph to describe the process of hyperparameter tuning "(3) Hyper-parameters optimization. All network configuration parameters (i.e., leaf number, number of trees, algorithm, and so on) were modified by a trial and error method to obtain the optimized network structure. The optimized RF model parameters are shown in Table S2. Figures S12 and S13 show the examples to optimize the number of min sample split and tree, respectively.".

Q9: Line 143, 145: please clarify how this "tiny data noise" is added.

Reply: Thank you. The "tiny data noise" was through randomly permutes the observations of feature *i*. We have revised it to "errOOB1 and errOOB2 represent the out-of-bag data error of feature *i* before and after randomly permuting the observation, respectively." in line 163-165 in the revised manuscript.

Q10: Line 147-150: I am not sure what the purpose is. Given the atmospheric lifetime of ozone, the measurements may well carry some "memory effects". Do the authors suggest that the ozone records can be reconstructed with randomly arranged inputs and hence any transport footprint (and factors alike) is already captured by the suite of instantaneous measurements?

Reply: Thank you. we agree with you that the measurements (especially O_3) may carry some "memory effects". However, it is inevitable that some "memory effects" cannot be considered. In order to decrease the dependence of the model on the data, we fed the randomly arranged inputs to the model by interrupting the continuity of the time series to check the robustness of the model. It is actually a cross-validation but with a smaller

fold (not systematic) than that used in the revised manuscript, in which we carried out a 12-fold cross-validation to split the training and testing data according to your suggestion. Therefore, the results are more robust in the revised manuscript. We deleted the sentence "we interrupted the continuity of the time series, fed the randomly arranged inputs to the model" in the revised manuscript, and added the detailed workflow in the Text S3.

Q11: Line 160: It is unclear how the initial VOC concentrations are derived. I do notice that Text S2 in the supplement information is about initial VOCs. Please refer to SI contents whenever necessary. I also have several major concerns regarding the initial VOC calculations in Text S2: (1) these formulations require initial/source ratios. It remains absolutely unclear how the authors derive the initial or source ratio of ethylbenzene/xylene. (2) The initial/source ratios may vary among different sources (e.g., gasoline, diesel, combustion, …). How would the authors account for the impacts from different emission sources? Or the mix with air masses with different photochemical ages? (3) Ethylbenzene and xylene are primarily of anthropogenic origin. The underlying assumption here is that all VOCs experience similar transport and aging processes. This may not be a bad assumption for other anthropogenic VOCs, but I doubt this could be applied for biogenic VOCs, which, according to the authors (Line 261 and Figure 3), is quite important. To sum up, I am not convinced that the initial VOC calculation presented in this manuscript can accurately describe the VOC oxidation during the transport receptor site.

Reply: Thanks for your comments. In the revised manuscript, we referred to SI contents accordingly, such as, "The calculation of initial VOCs and sensitivity tests can be found in the Supplemental Materials (S2)" in lines 142-143 and 190-191.

We took the mean value (0.52) of ethylbenzene/xylene from 0.45 to 0.66 before sunrise during our observations as the initial ratio of ethylbenzene/xylene. The protocol is the same as that reported in literatures (Shao et al., 2011; Zhan et al., 2021).

We agree with you that the initial/source ratios may vary among different sources (e.g., gasoline, diesel, combustion). Here, we used the ambient concentrations of

ethylbenzene and xylene to calculate the initial source ratio. Thus, it actually reflects a mean or mixing value of different sources instead of a specific one. There is a basic assumption that emission patterns of different sources are relatively stable in a day when calculating the photochemical initial concentrations of VOCs using this method (Yuan et al., 2013). As replied in the third question (Q3), we checked this assumption based on PSCF analysis. The results suggest that the emission patterns of xylene and ethylbenzene are highly similar in the early morning to that in the whole day.

Although xylene and ethylbenzene are anthropogenic VOCs, while isoprene is a biogenic VOC, PSCF analysis indicated that the spatial pattern of isoprene was similar to that of xylene and ethylbenzene during our observations (Figure R8). This means VOC emissions were evenly distributed in Beijing. This can be ascribed to the fact that our observation site is a typical urban station and VOCs are emitted from non-point sources on a city scale. Therefore, the photochemical clock calculated using xylene and ethylbenzene is able to correct the photochemical loss of biogenic VOCs to some extent.

In Text S2 in the revised SI, we added more details to calculate the initial concentrations of VOCs.

"**Text S2. Calculation of initial VOCs concentrations**

Photochemical initial concentration (PIC) proposed by Shao et al. (2011), which is calculated based on the photochemical-age approach and has been applied to evaluate the effect of photochemical processing on measured VOC levels. Equation S1 essentially describes the integrated OH exposure (Shao et al., 2011).

$$\int c_{OH}dt = \frac{1}{k_{A,OH}-k_{B,OH}}\left[\ln\left(\frac{VOC_A}{VOC_B}\right)_{initial} - \ln\left(\frac{VOC_A}{VOC_B}\right)\right] \quad (S1)$$

The initial concentration of species *i* can be calculated using Equation S2.

$$VOC_{i,\,initial} = \frac{VOC_i}{\exp(-k_{i,OH})\exp(\int c_{OH}dt)} \quad (S2)$$

Substituting equation 1 into equation 2, then we can get equation S3.

$$VOC_{i,\,initial} = \frac{VOC_i}{\exp(-k_{i,\,OH})\exp\left(\frac{1}{k_{A,OH}-k_{B,\,OH}}\left[\ln\left(\frac{VOC_A}{VOC_B}\right)_{initial} - \ln\left(\frac{VOC_A}{VOC_B}\right)\right]\right)} \quad (S3)$$

Where C_{OH} represents the ambient OH concentration; $k_{A,OH}$ and $k_{B,OH}$ represent the reaction rate of compound A and B with OH radical, respectively; t represents the reaction time of species *i* in the ambient.

In previous work (Shao et al., 2011; Zhan et al., 2021), the selection of ethylbenzene and xylene as tracers was justified for calculating ambient OH exposure under the following conditions: 1) the concentrations of xylene and ethylbenzene were well correlated (Figure S9), which indicated that they were simultaneously emitted; 2) they had different degradation rates in the atmosphere; and 3) the calculated PICs were in good agreement with those calculated using other tracers (Shao et al., 2011; Zhan et al., 2021).

In this study, the ethylbenzene/xylene pair was used to calculate ambient OH exposure. As shown in Figure S9, the concentrations of xylene and ethylbenzene are well correlated, which indicates that they are simultaneously emitted. In addition, we compared the PICs according to xylene/ethylbenzene with that using toluene/benzene (Figure S10). The calculated PICs ratio ($PIC_{Xylene/Ethylbenzene}$ / $PIC_{Toluene/Benzene}$) varied from 0.5 to 1.5 with a mean value of 0.96. This means the calculated initial VOCs was in good agreement when using different tracers. The mean ratio (0.52, from 0.45 to 0.66) of ethylbenzene/xylene before sunrise was taken as the initial ratio of ethylbenzene/xylene. Sensitivity tests showed that the uncertainty of PICs caused by the OH exposure (from −10% to +10%) ranged from 0.55 to 1.57 (Table S4).

Variations of air mass may also affect the VOC ratio. Figure S11 A-D shows the mean concentration distribution of ethylbenzene and xylene in the early morning and the whole day based on potential source contribution function (PSCF) analysis. Xylene showed similar patterns to ethylbenzene in different air mass trajectories and different periods. These results indicate that the emissions of xylene and ethylbenzene were constant throughout the day and variations of air mass should have little influence on the initial ratio of VOCs. The hourly concentrations of ethylbenzene and xylene were used to calculate the concentration of initial VOCs. The initial VOC was calculated by adding the measured VOC concentration and the calculated photochemical loss. Figure S12 shows the diurnal variations of the observed and initial VOCs concentrations from 2014 to 2016. Photochemical loss of VOC occurred mainly during the daytime.

It should be noted that the lifetimes ($1/k_{2COH}$) of highly reactive VOCs, such as isoprene, greatly depend on the OH exposure. The photochemical ages of isoprene were

0.01–6.21 h (1.26 ± 1.12 h). This value is comparable with previously reported photochemical ages (Shao et al., 2011; Gao et al., 2018). However, the initial concentrations of highly reactive VOCs may be overestimated due to their short lifetimes and should be taken as the upper limits. On the other hand, isoprene is a biogenic VOC, while xylene and ethylbenzene are anthropogenic VOCs. If they do not share the same air mass histories, an additional uncertainty is inevitable for the PICs of isoprene. As shown in Figure S11, isoprene showed similar patterns to that of xylene and ethylbenzene, which means VOC emissions are evenly distributed in Beijing during our observations. This can be ascribed to the fact that our observation site is a typical urban station. Although isoprene and xylene/ethylbenzene different sources, both them are non-point sources on a city scale. Therefore, the photochemical clock calculated using xylene and ethylbenzene is able to correct the photochemical loss of biogenic VOCs to some extent. It should be noted that uncertainty is inevitable when we estimating the photochemical age (Parrish et al., 2007). However, the aim of this work is to test whether the ML-model can reflect the influence of photochemical loss of VOCs species on O_3 modelling. The PICs should provide additional information for understanding O_3 formation in the atmosphere.”.

Q12: Line 176-177: Simply comparing RF outputs to measurements ABSOLUTELY DOES NOT guarantee that the RF model is robust. An over-trained model will also show good performance when evaluated with measurements. Figure 2 says nothing about the robustness of the model.

Reply: Thank you so much. We agree with you that simply comparing RF outputs to measurements absolutely does not guarantee that the RF model is robust. We have carried out cross-validation in the revised manuscript according to your good suggestion. The details on the robustness of the model have been replied in Q1 and Q2. All the figures (Figure 2, Figure S3 and Figure S4) have been updated in the revised manuscript and the revised SI.

Q13: Line 221-222: The authors stated that “ML is a black-box model” then cited

Sayeed et al. (2021). Do the authors then imply that the convolutional neural network model developed by and described in Sayeed et al is a "black-box"? If this is the case, please elaborate how the Sayeed et al model appears to be a black-box to the authors, since this is a somewhat strong accusation. I also do not fully agree with this statement. Random forest is actually fairly transparent compared to some other types of machine learning algorithms, as one can certainly examine the trees and see how a certain feature is used for splitting the nodes and how the overall importance is propagated, if they wish.

Reply: Thank you. We agree with you that when compared with other types of machine learning algorithms, random forest is actually fairly transparent because one can examine every node. However, when compared with chemical models, such as box model and chemical transport model. It is still a "black box" from the point view of chemical mechanism. In lines 116-118 in the revised manuscript, we revised this sentence to "Although ML is widely used to understand air pollution, explanations of ML results (e.g., RI) are somewhat vague because ML is a black-box model from the point view of chemical mechanism".

Q14: Line 267-270, Line 287-293: Please clarify how the EKMA plots were generated using the RF model. It is confusing that Figures 4A-C and Figure 4D use different color scales. It also appears to me that Figure 4B and Figure 4D show considerable discrepancy: the "observed point" in Figure 4B indicates ~44 ppb ozone but in Figure 4D the ozone level at the "observed point" is ~60 ppb. Please also define this "observed point". It could be that the OBM predicts some sort of "maximum ozone production potential" driven by chemistry and impacts like transport would not be captured. But this would be inconsistent with one of the conclusions of this work (e.g., Line 206) that ozone production seems to be dominated by local chemistry. Either way, the discrepancy needs to be elaborated.

Reply: Thank you for your good comment. In the revised SI, we have added the calculation process of EKMA in Text S3. This has also been replied in **Q2**. In lines 171-179 in the revised manuscript, we added a paragraph "When plotting the O_3 formation

sensitivity curves, we made a virtual matrix of inputs by varying the concentrations of NOx and VOCs from 0.9 to 1.1 times (with a step of 0.01) of their mean values while keeping all other inputs unchanged (i.e., the mean values). Then, the new matrix was used as testing data, while all the measured data were taken as training data. Thus, the testing data should represent the mean sensitivity regime of O_3 in Beijing, while the training data actually covered all the sensitivity regimes of O_3 formation. The EKMA curves were plotted using the daily maximum 8-h (MDA8) O_3. More details can be found in the SI".

In the initial submission version of the manuscript, the EKMA curves based on the RF model were plotted using the mean O_3 concentrations in a day with different color scale, while the EKMA curves from the Box model were plotted using the maximum O_3 concentrations in a day. This leads to the difference in O_3 concentrations between Figure 4B and 4D. In the revised manuscript, we plotted the EKMA curves from both models using the daily maximum 8-h (MDA8) O_3 with the same scale. As shown in Figure R3 or Figure 4. The predicted O_3 concentrations using two different models are comparable. In addition, we calculated the relative errors (within 15.6%) of O_3 concentrations from the RF model compared with that of the Box model (Figure R4 or Figure S8). The observed point has been added in the revised Figure 4 or Figure R3.

In Text S3 in the revised SI, we have added the details for plotting EKMA curves: "(6) EKMA curves. The Empirical Kinetic Modeling Approach (EKMA) was used to assess the O_3 formation mechanism regime. Both the RF model and a box model with Master Chemical Mechanism (MCM, 3.3.1) were used to calculate the EKMA curves. For the RF model simulations, the observed point data was chosen as the mean values of the input parameters during our observations, then the concentrations of VOCs and NOx were varied ±10% (or from 90% to 110%) of their mean values with a step of 1% in a two-dimensional matrix along with other inputs unchanged. This matrix was used as the testing data, while all the measured data were taken as the training data in the RF model to simulate O_3 concentrations under different scenarios of VOCs and NOx concentrations. To decrease the model uncertainty, we set relatively small variations of VOCs and NOx (±10%) compared to the observed values in this study. The mean

relative error of simulated O_3 concentrations between RF model and Box model (within 15.6%, Figure S8) suggests that the RF model can well predict O_3 concentrations during our observations.”.

Q15: Line 299-301: I actually fail to see this. Figures 2E and 2F show that the improvement is small.

Reply: Thank you. In the revised manuscript, we have revised the sentence “Thus, our model has good prediction performance (R^2 = 0.87) when combined with the initial VOC species…and identify the connection between the reactivity of VOC species and O_3 formation in the atmosphere” to “When the TVOCs were split into measured or initial VOC species, the R^2 increased obviously as the number of data features increased. Therefore, the VOC composition has a significant influence on O_3 prediction using the RF model.” in Lines 221-223 in the revised manuscript. Meanwhile, we added “To verify the stability of the model, we performed a significance test on the model results. The results showed that there was no significant difference among the different tests (P>0.05, R^2>0.98).” in Lines 168-170 in the revised manuscript.

Minor/technical comments:

Q16: Line 16: Beijing, China.

Reply: Thank you. It has been corrected in Line 14 in the revised manuscript.

Q17: Line 21: Define abbreviations upon first appearance. NOx and $PM_{2.5}$ are not defined. Also, x should be subscripted throughout.

Reply: Thank you. We have defined the abbreviations when it first appears, such as “nitrogen oxides (NO_x) and relative humidity (RH), and a positive response to temperature (T), solar radiation (SR)” and “fine particulate matter with aerodynamic diameter less than 2.5 μm ($PM_{2.5}$)” in Lines 22 and 95-96. The “x” in NO_x has been subscripted throughout the paper.

Q18: Line 22: Delete curves

Reply: Thank you. We have deleted the "curves"

Q19: Line 33: "NO-O_3-NO_2 cycle" please elaborate. The audience will appreciate clearly described concepts.

Reply: Thank you. The "NO-O_3-NO_2 cycle" can be expressed by eq. R (1) and R (2).

$$HO_2/RO_2 + NO \rightarrow OH/RO + NO_2 \qquad R\ (1)$$

$$NO_2 + O_2 \xrightarrow{hv} NO + O_3 \qquad R\ (2)$$

We have revised "NO-O_3-NO_2 cycle" to "conversion from NO to NO_2, subsequently, formation of O_3 by photolysis of NO_2 in the presence of O_2" in Line 36 in the revised manuscript.

Q20: Line 31-34: Very busy and ill-formed sentence. Please rephrase.

Reply: Thank you. We have revised this sentence to "Oxidation of volatile organic compounds (VOCs) will produce peroxyl radicals (RO_2) and hydroperoxyl radicals (HO_2). The RO_2/HO_2 can accelerate conversion from NO to NO_2, subsequently, formation of O_3 by photolysis of NO_2 in the presence of O_2 (Wang et al., 2017a) " in lines 34-37 in the revised manuscript.

Q21: Line 52-53: "In addition, this method lacks the predictability of O_3 concentrations for policy-making." This is a bizarre statement. By definition, observations do not have predictability.

Reply: Thank you. We have deleted this statement in the revised manuscript.

Q22: Line 53: OBM can use remote sensing measurements too.

Reply: Thank you. We have revised to "OBMs combine *in-situ* field observations, remote sensing measurements and chemical box models, which…" in Line 56 in the revised manuscript.

Q23: Line 80: has -> have.

Reply: Thank you. We have revised "has" to "have" in Line 88 in the revised

manuscript.

Q24: Line 99: please fix citation style as per the journal requirements.

Reply: Thank you. We have fixed citation style as per the journal requirements in the revised manuscript.

Q25: Line 98-100: "Makar et al (1999) reported that highly reactive species, such as isoprene, were underestimated by 40% when the OH reactions were ignored" This is confusing as currently written, which is also a misinterpretation of the paper. Makar et al. clearly stated that it is the isoprene emissions that are underestimated if the OH oxidation is not considered.

Reply: Thank you. We have revised it to "the isoprene emissions were underestimated by up to 40% if the OH oxidation is not considered." in Lines 109-110 in the revised manuscript.

Q26: Line 101: Please clearly define what exactly "initial concentration" is. It is my opinion that this concept is vague: if one attempts to measure this "initial concentration", how close shall the sensor be placed? When it comes to biogenic compounds like isoprene, shall the sensor be placed at the leaf level, within the canopy, or what?

Reply: Thank you. The "initial concentration" is a concept based on reaction time instead of location or space. It is defined as the concentrations of VOCs before degradation has occurred in a day and actually extrapolated according to a "chemical clock" (OH exposure).

In lines 118-121 in the revised manuscript, we clarified it "In this study, we used the RF model to evaluate the prediction performance of atmospheric O_3 using the TVOCs, measured VOC species and photochemical initial concentration (PIC) of VOC species, which is calculated based on the photochemical-age approach (Shao et al., 2011)" and in the revised SI, we added a sentence "Photochemical initial concentration (PIC) proposed by Shao et al. (2011), which is calculated based on the photochemical-age approach and has been applied to evaluate the effect of photochemical processing

on measured VOC levels".

Q27: Line 133: Random forest is not a type of decision tree. It is a collection of a number of decision trees.

Reply: Thank you. We have revised it to "The random forest (RF) is a type of ensemble decision tree that can be used for classification and regression" in Line 145 in the revised manuscript.

Q28: Line 132-: this section does not mention anything about how these input variables are pre-processed: are extreme values removed? Do the authors apply any standardization or normalization? These variables are on very different numerical scales and sometimes standardization could improve the stability and/or performance.

Reply: Thank you. We carried out data-normalization and removed the extreme values before modelling. In the revised SI, we clarified it and added the data processing in Text S3 "(2) Data process. After the extreme values were removed, all data were normalized (between 0 and 1) in order to decrease the sample distribution range, accelerate calculation efficiency and improve the robustness of the RF model. Then, the dataset was randomly divided into 12 subsets. Thus, a 12-fold cross-validation was performed by alternately taking one subset as testing data and the rest as training data to ensure that each data point has an equal chance being trained and tested.".

Q29: Line 136: Please define "score".

Reply: Thank you. To describe the random forest more accurately, we have revised "score" to "result" in line 146-148 in the revised manuscript.

"During the training process, the model creates a large number of different decision trees with different sample sets at each node, and then averages the results of all decision trees as its final results".

Q30: Figure 1: what are those red arrows in the top panel?

Reply: Thank you. The red arrows represent the O_3 concentration exceeding 74.6 ppbv

according to the national ambient air quality standard. We have added this explanation in Figure 1 or Figure R11 in Lines 205-206 in the revised manuscript.

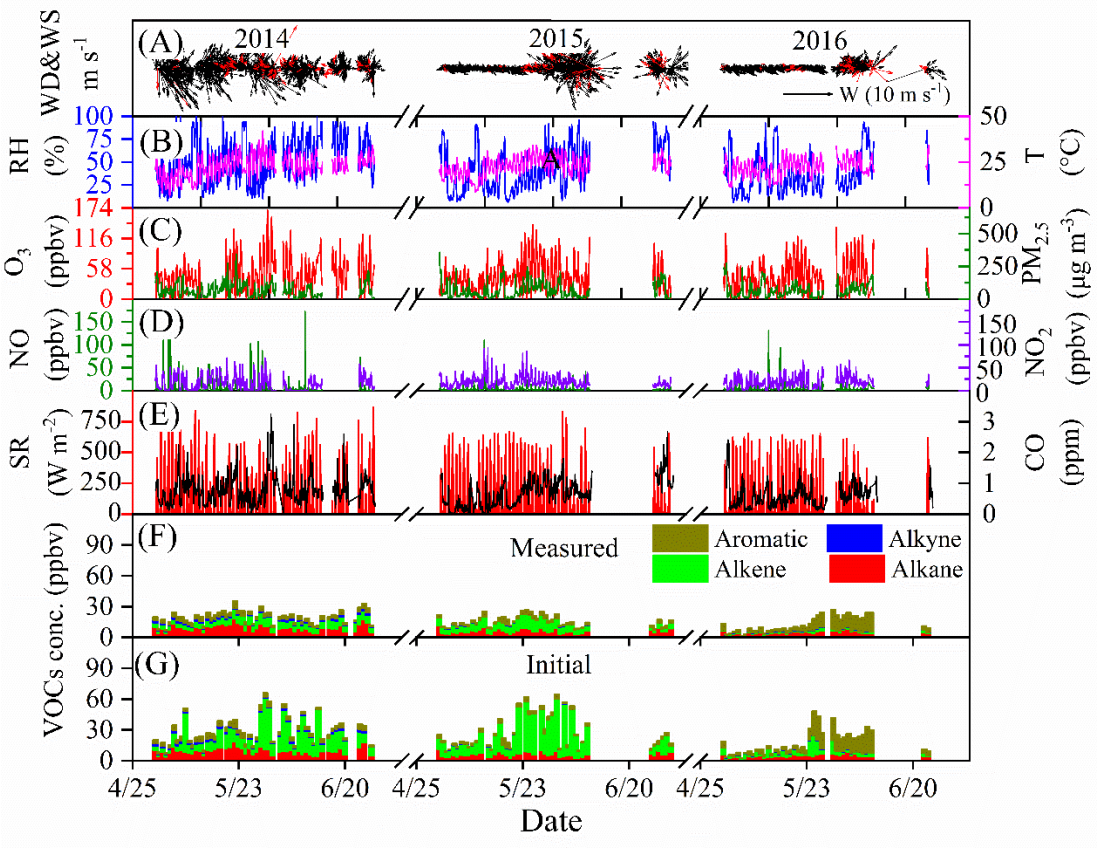

Figure R11. Time series of air pollutants and meteorological parameters during observations in Beijing. (In A, the red arrows represent the O_3 concentration exceed 74.6 ppbv according to the national ambient air quality standard.)

Q31: Figure 1: it would be interesting to separate biogenic VOCs from anthropogenic VOCs.

Reply: Thank you. In this figure, we want to show the general characteristics of different kinds of VOCs. It is usually classified according to the structures of VOCs. Figure R12 shows the time series of biogenic and anthropogenic VOCs. We also added this figure in the revised SI (Figure S16).

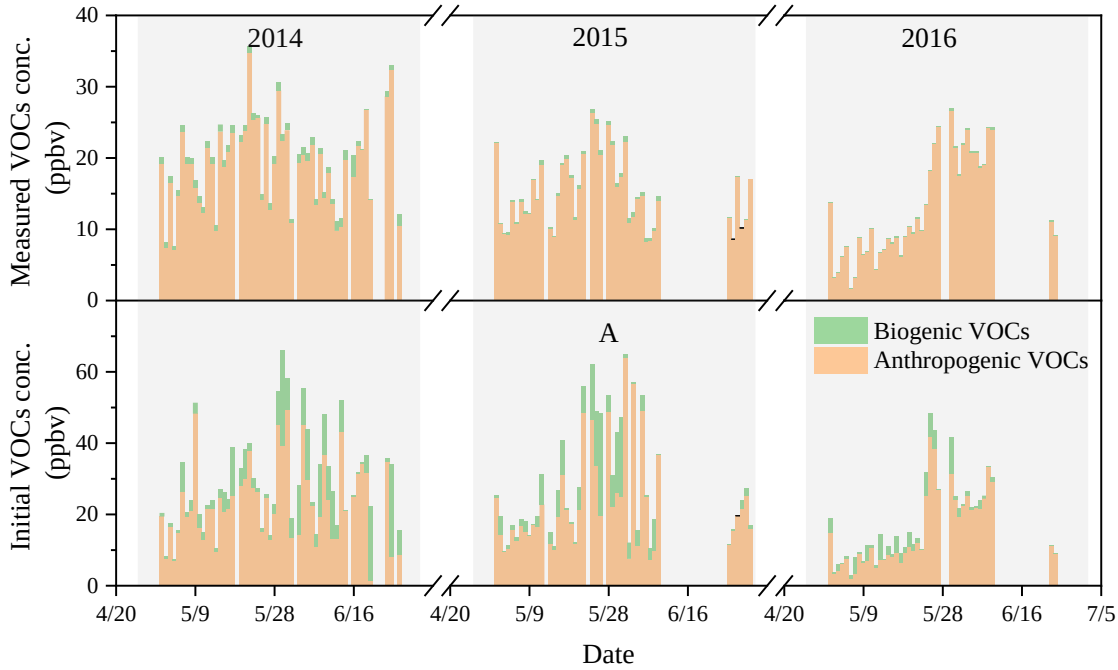

Figure R12. Time series of biogenic and anthropogenic VOCs during the observation period. (Biogenic VOCs: including isoprene, α-pinene and β-pinene; Anthropogenic VOCs: including all detected VOCs except isoprene, α-pinene and β-pinene)

Q32: Figure 2: all top panels are not readable. Please consider extending the time series plots.

Reply: Thank you. We have revised Figure 2 (or Figure R2) in the revised manuscript.

Q33: Text S1: "total of 51 VOCs (including 21 alkanes, 13 alkenes, 1 alkyne and 16 aromatics) were analyzed within a limit of quantification of 0.1-100 ppbv". Well this is a very wide range for limit of quantification. Please include a table and list the limits of quantification/detection for all VOCs used in this work. I'd argue that if the LOQ/LOD of an ambient VOC is on the order of 100 ppb, the data is practically useless. The total VOC levels are less than ~30 ppb (Figure 1F).

Reply: Thank you. In the submission version, we made a mistake about the number. It was 0.002-0.05 ppbv instead of 0.1-100 ppbv. We have added the method detection limit (MDL) for all measured species in Table R2 or Table S4. In Text S1, we have

revised "total of 51 VOCs (including 21 alkanes, 13 alkenes, 1 alkyne and 16 aromatics) were analyzed within a limit of quantification of 0.1-100 ppbv" to "total of 51 VOCs (including 21 alkanes, 13 alkenes, 1 alkyne and 16 aromatics) were analyzed within a limit of quantification of 0.002-0.05 ppbv as shown in Table S4".

Q34: Text S2: please define all variables. C_{OH}, t, k, are not defined. Also, Equation S1 essentially describes the integrated OH exposure, rather than the "changes in VOC concentration as a function of time due to photochemical reaction".

Reply: Thank you. We have defined the all variables and revised the description "Equation S1 essentially describes the integrated OH exposure (Shao et al., 2011)" in Text S2 in the revised SI.

"Where C_{OH} represents the ambient OH concentration; $k_{A,OH}$ and $k_{B,OH}$ represent the reaction rate of compound A and B with OH radical, respectively; t represents the reaction time of species *i* in the ambient."

Q35: Table S1: please define all parameters. What's leaf number? Is this the number of leaf nodes, leaf levels/depths, or what? What's fboot?

Reply: Thank you, we have revised the Table S2 (Table R3). The leaf number represents the minimum number of samples to be split. The fboot represents the minimum number of samples in a leaf. They are added in the footnotes of Table S2.

Reference

Breiman, L. Random Forests. Machine Learning, 45, 5-32, 10.1023/A:1010933404324, 2001.

Carter, W.P.L. Development of the SAPRC-07 chemical mechanism. Atmospheric Environment, 44, 5324-5335, https://doi.org/10.1016/j.atmosenv.2010.01.026, 2010.

Gao, J., Zhang, J., Li, H., Li, L., Xu, L., Zhang, Y., Wang, Z., Wang, X., Zhang, W., Chen, Y., Cheng, X., Zhang, H., Peng, L., Chai, F., Wei, Y. Comparative study of volatile organic compounds in ambient air using observed mixing ratios and initial mixing ratios taking chemical loss into account – A case study in a typical urban area in Beijing. Science of The Total Environment, 628-629, 791-804, https://doi.org/10.1016/j.scitotenv.2018.01.175, 2018.

Hou, L., Dai, Q., Song, C., Liu, B., Guo, F., Dai, T., Li, L., Liu, B., Bi, X., Zhang, Y., Feng, Y. Revealing Drivers of Haze Pollution by Explainable Machine Learning. Environmental Science & Technology Letters, 10.1021/acs.estlett.1c00865, 2022.

Liu, H., Liu, J., Liu, Y., Ouyang, B., Xiang, S., Yi, K., Tao, S. Analysis of wintertime O3 variability using

a random forest model and high-frequency observations in Zhangjiakou—an area with background pollution level of the North China Plain. Environmental Pollution, 262, 114191, https://doi.org/10.1016/j.envpol.2020.114191, 2020.

Liu, Z., Qi, Z., Ni, X., Dong, M., Ma, M., Xue, W., Zhang, Q., Wang, J. How to apply O3 and PM2.5 collaborative control to practical management in China: A study based on meta-analysis and machine learning. Science of The Total Environment, 772, 145392, https://doi.org/10.1016/j.scitotenv.2021.145392, 2021.

Ma, R., Ban, J., Wang, Q., Zhang, Y., Yang, Y., He, M.Z., Li, S., Shi, W., Li, T. Random forest model based fine scale spatiotemporal O3 trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017. Environmental Pollution, 276, 116635, https://doi.org/10.1016/j.envpol.2021.116635, 2021.

Parrish, D.D., Stohl, A., Forster, C., Atlas, E.L., Blake, D.R., Goldan, P.D., Kuster, W.C., de Gouw, J.A. Effects of mixing on evolution of hydrocarbon ratios in the troposphere. Journal of Geophysical Research: Atmospheres, 112, https://doi.org/10.1029/2006JD007583, 2007.

Shao, M., Bin, W., Sihua, L., Bin, Y., Ming, W. Effects of Beijing Olympics Control Measures on Reducing Reactive Hydrocarbon Species. Environmental Science & Technology, 45, 514-519, 10.1021/es102357t, 2011.

Wang, T., Xue, L., Brimblecombe, P., Lam, Y.F., Li, L., Zhang, L. Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects. Science of The Total Environment, 575, 1582-1596, https://doi.org/10.1016/j.scitotenv.2016.10.081, 2017a.

Wang, Y., Li, Y., Pu, W., Wen, K., Shugart, Y.Y., Xiong, M., Jin, L. Random Bits Forest: a Strong Classifier/Regressor for Big Data. Scientific Reports, 6, 30086, 10.1038/srep30086, 2016.

Wang, Y., Wu, G., Deng, L., Tang, Z., Wang, K., Sun, W., Shangguan, Z. Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm. Scientific Reports, 7, 6940, 10.1038/s41598-017-07197-6, 2017b.

Yuan, B., Hu, W.W., Shao, M., Wang, M., Chen, W.T., Lu, S.H., Zeng, L.M., Hu, M. VOC emissions, evolutions and contributions to SOA formation at a receptor site in eastern China. Atmos. Chem. Phys., 13, 8815-8832, 10.5194/acp-13-8815-2013, 2013.