Reviewer 1#

In this revised manuscript, the authors addressed most of my previous concerns. Specifically, the revised manuscript now includes more details about the random forest model as well as the relevant measures to mitigate overtraining; the authors also discussed in greater detail the initial VOCs. I appreciate the authors' efforts and it is my opinion that the quality of the manuscript is greatly improved. I recommend the manuscript for publication after the following minor comments are addressed, which are intended to further improve the clarity and the flow:

Response: Thank you for your good comments and suggestions. 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.

Line 43: … exhibiting

Response: Thank you. We have revised it to "exhibiting" in Line 43.

Line 41-53: consider combining this paragraph with the next for a smoother flow.

Response: Thank you. We have combined these two paragraphs to make it smoother.

"…Therefore, it is urgent to develop an accurate and highly efficient method for timely assessing the sensitivity regime of O_3 production and evaluating the effectiveness of a potential measure on O_3 pollution control. The sensitivity of O_3 formation can usually be analysed using observed indicators…"

Line 99-101: This sentence touches one common concern that some machine learning models may be less transparent/interpretable compared to other conventional techniques, thus giving the impression that the authors would discuss this concern in greater detail in this paragraph. Yet, the rest of this paragraph drifted away. Then the next paragraph opens with yet another statement on this "black box" concern. Please rewrite these two paragraphs to improve the logic and flow.

Response: Thank you. We have rewritten these two paragraphs to improve the logic and flow.

"Although ML is widely used to understand air pollution, many ML studies have

used total VOCs (TVOCs) to simulate O_3 formation and rarely considered the effect of VOC species on O_3 formation sensitivity (Feng et al., 2019; Liu et al., 2021; Ma et al., 2021a). Thus, they were unable to identify the chemical reactivity of a single species to O_3 formation, which may lead to underestimations or even misunderstandings of the role of VOCs in O_3 formation because the same concentration of TVOCs with different compositions may lead to different OPEs. In addition, VOCs react with OH radicals during atmospheric transport, which is the most important sink of VOCs (Carlo et al., 2004; Liu et al., 2020b). Makar et al. (1999) reported that the isoprene emissions were underestimated by up to 40% if the OH oxidation is not considered. Other studies indicated that the initial concentrations of VOCs, which account for the photochemical loss of VOCs during transport, were more representative of pollution levels in the sampling area than the observed VOCs (Yuan et al., 2013; Zhan et al., 2021). However, whether the ML model can identify the connection between the reactivity of VOC species and O_3 formation sensitivity has not been clarified.

It should be noted that physical interpretability of the results is an important question when ML models are applied in atmospheric studies (Hou et al., 2022). However, explanations of ML results (e.g., RI) are somewhat vague because ML is a "black-box" model from the point view of chemical mechanism (Hou et al., 2022; Taoufik et al., 2022)…."

Line 235-237: This is a healthy start, but the outcome of this 12-fold cross validation is missing. Please include a figure or table, perhaps in the SI, to archive the consistency of the model performance across all 12 folds. Whether splitting the dataset randomly is a good strategy for cross validation remains a subject of debate, but it is key to archive all key details.

Response: Thank you. We have added the results of 12-fold cross validation in **Table R1** or **S6** and performed Pearson correlation calculations as shown in **Table R2** or **S7**. It can be seen in Table R2 that there is consistency in the results between the different folds (ranged from 0.95 to 0.98).

61 **Table R1**. The RI values of 2015 in all 12 folds

species number	Fold number											
	1	2	3	4	5	6	7	8	9	10	11	12
Ethane	1.39	1.15	1.69	2.28	1.70	1.70	1.97	1.80	1.20	1.35	1.71	1.26
ethene	1.60	1.97	1.01	1.01	1.52	2.15	0.65	2.01	1.44	2.18	2.17	1.64
Propane	0.24	0.85	0.38	1.20	1.23	1.18	0.49	0.01	1.19	0.54	0.28	1.61
propene	5.70	5.95	6.30	5.88	6.06	5.14	6.59	6.18	5.94	5.99	6.10	5.28
iso-Butane	1.45	0.96	1.58	1.55	1.67	1.22	1.77	1.97	1.26	1.80	1.64	1.56
n-Butane	0.05	1.08	0.35	0.32	0.35	0.61	0.27	0.24	0.60	0.05	0.09	0.05
Acetylene	1.50	1.11	1.01	1.03	0.66	0.00	0.96	0.96	0.23	1.05	1.02	1.81
trans-2-Butene	1.30	2.26	2.37	2.09	1.51	2.22	2.23	2.30	1.93	2.42	2.60	2.34
1-Butene	1.71	1.29	0.89	1.13	1.20	1.34	1.31	1.17	1.70	1.20	0.93	1.42
Cyclopentane	0.00	0.00	0.25	0.00	0.25	0.25	0.24	0.24	0.27	0.00	0.25	0.24
cis-2-Butene	0.86	1.05	1.33	1.09	1.43	1.65	1.65	1.23	1.60	0.81	1.44	1.15
iso-Pentane	0.00	0.00	0.30	0.00	0.25	0.25	0.25	0.24	0.23	0.00	0.23	0.27
1,3-Butadiene	0.58	1.31	0.86	1.10	0.85	1.17	0.69	0.99	1.36	0.91	1.08	0.93
trans-2-Pentene	0.99	0.75	1.26	0.78	1.16	1.18	1.03	1.01	0.69	1.17	1.12	1.23
1-Pentene	0.28	0.70	0.81	0.77	0.59	1.06	0.61	0.92	0.86	0.56	1.07	0.49
cis-2-Pentene	1.29	1.37	1.16	1.21	1.26	0.84	1.23	1.76	1.55	1.09	1.25	0.98
2,3-Dimethylbutane	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.26	0.00	0.00
n-hexane	1.27	1.42	1.69	1.46	1.42	1.47	1.30	1.62	0.82	2.00	1.08	0.91
isoprene	5.30	4.16	4.96	5.25	4.98	5.39	5.98	5.25	5.44	5.06	5.34	6.24
2-methyl-1-pentene	2.27	2.51	2.08	1.93	2.36	2.09	2.10	1.74	2.35	2.28	2.26	2.17
2,2-Dimethylbutane	0.78	0.63	0.65	0.68	0.87	0.63	0.41	0.57	0.81	0.71	0.63	0.60
Benzene	0.35	0.79	0.96	1.37	0.84	1.12	0.86	0.83	0.80	1.10	1.10	0.79
cyclohexane	1.71	1.40	1.18	1.20	1.46	1.06	1.55	1.09	0.86	1.21	1.22	1.61

2,3-Dimethylpentane	2.26	1.61	2.39	1.65	2.23	1.93	1.68	1.56	1.89	2.16	2.01	2.21
3-Methylhexane	1.05	0.36	0.63	0.98	0.48	0.92	0.84	0.69	0.54	0.45	0.90	0.66
2,2,4-Trimethylpentane	1.59	1.64	1.62	1.72	1.28	1.33	1.46	1.59	1.85	2.07	1.84	1.71
n-heptane	1.19	1.13	1.15	1.16	1.36	0.97	0.95	0.60	1.06	1.27	1.17	0.83
methylcyclohexane	0.34	0.56	0.97	0.69	0.50	0.55	0.75	0.42	0.57	0.70	0.81	0.86
2,3,4-trimethylpentane	0.76	0.76	0.87	1.11	0.89	0.90	0.81	0.67	0.95	0.81	0.88	0.97
Toluene	0.79	0.47	0.63	0.50	0.75	0.39	0.49	0.64	0.91	0.71	0.68	0.52
2-methylhexane	1.56	1.69	1.39	1.32	1.61	1.73	1.37	1.50	1.02	1.93	1.08	1.25
3-methylhexane	0.47	0.31	0.51	0.83	0.79	0.70	0.54	0.78	0.47	0.50	0.67	0.61
n-octane	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ethylbenzene	0.73	0.81	1.10	0.83	0.48	0.72	0.51	0.61	0.68	0.79	0.79	0.92
m,p-Xylene	0.52	0.31	0.83	0.22	0.32	0.24	0.45	0.36	0.01	0.53	0.57	0.22
styrene	0.31	0.41	0.23	0.27	0.25	0.26	0.33	0.33	0.37	0.17	0.29	0.28
o-xylene	0.59	0.83	0.88	0.61	0.62	0.98	0.63	0.64	0.86	0.83	0.66	0.30
n-nonane	1.19	1.30	1.38	1.68	1.34	1.57	1.30	1.14	1.48	1.27	1.27	1.73
iso-Propylbenzene	0.33	0.13	0.49	0.32	0.51	0.25	0.29	0.11	0.01	0.34	0.39	0.22
a-pinene	2.10	1.47	1.67	2.02	1.87	1.97	1.77	2.43	1.53	1.72	1.60	1.83
n-Propylbenzene	2.03	1.91	1.88	2.15	1.89	2.24	1.69	1.86	1.69	1.88	1.91	2.23
m-ethyl toluene	1.54	0.95	0.91	1.08	0.79	0.90	0.64	1.02	0.99	0.97	0.64	0.95
p-ethyltoluene	0.55	0.64	1.22	0.74	1.06	0.98	0.74	1.01	0.73	0.84	0.59	0.53
1,3,5-Trimethylbenzene	0.80	1.30	0.80	1.06	0.78	1.01	1.35	0.97	1.18	1.18	0.86	1.00
o-ethyl toluene	1.00	0.98	0.98	0.73	0.71	0.91	1.09	0.59	0.84	0.60	0.88	1.03
β-pinene	0.20	0.06	0.23	0.01	0.25	0.14	0.02	0.46	0.05	0.04	0.25	0.01
1,2,4-trimethyl benzene	1.38	1.67	1.54	1.65	2.69	1.34	1.69	1.56	1.03	1.43	1.60	0.91
n-decane	1.60	1.84	2.02	0.93	1.52	1.88	1.91	1.58	1.97	1.53	1.93	1.58
1,2,3-trimethyl benzene	1.14	1.05	1.80	1.22	0.96	1.70	1.37	1.11	0.67	0.88	1.53	1.32

m-diethyl benzene	0.26	0.00	0.25	0.26	0.24	0.36	0.00	0.34	0.00	0.26	0.00	0.00
p-diethyl benzene	1.19	0.95	0.95	0.69	0.92	0.80	0.79	0.92	0.80	0.97	0.75	0.86
NO_x	16.35	16.32	14.22	15.74	15.93	11.94	14.37	14.00	15.79	14.78	13.75	14.23
T	10.25	9.69	9.45	9.69	9.18	10.77	10.61	10.74	9.81	8.76	8.71	9.50
RH	3.93	4.27	3.36	4.31	3.85	3.68	3.63	3.62	3.96	3.61	3.11	4.10
SR	3.35	3.64	3.94	3.86	3.34	3.73	3.77	3.72	3.44	3.30	4.25	3.83
WS&WD	4.28	3.08	4.29	3.15	2.63	3.59	3.42	4.05	4.13	3.52	4.63	3.71
$PM_{2.5}$	1.33	1.23	1.00	1.24	2.46	1.78	1.81	1.03	2.20	2.99	1.79	2.00
CO	3.00	4.03	2.97	2.72	2.91	3.88	3.78	3.64	3.41	3.05	3.54	2.49

Note: To verify the consistency between the different folds, we calculated the Pearson correlation coefficient (*r*) between the different folds as shown in **Table R2**.

Table R2. The Pearson correlation coefficient (*r*) between different folds in 2015

Fold number	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	0.98	0.98	0.98	0.98	0.95	0.98	0.98	0.98	0.98	0.97	0.98
2	0.98	1.00	0.98	0.98	0.97	0.95	0.97	0.97	0.98	0.98	0.96	0.96
3	0.98	0.98	1.00	0.98	0.97	0.96	0.98	0.98	0.98	0.98	0.98	0.97
4	0.98	0.98	0.98	1.00	0.98	0.96	0.98	0.98	0.98	0.97	0.97	0.98
5	0.98	0.97	0.97	0.98	1.00	0.96	0.98	0.97	0.98	0.98	0.97	0.97
6	0.95	0.95	0.96	0.96	0.96	1.00	0.98	0.98	0.96	0.95	0.97	0.96
7	0.98	0.97	0.98	0.98	0.98	0.98	1.00	0.98	0.98	0.97	0.98	0.98
8	0.98	0.97	0.98	0.98	0.97	0.98	0.98	1.00	0.98	0.97	0.98	0.97
9	0.98	0.98	0.98	0.98	0.98	0.96	0.98	0.98	1.00	0.98	0.97	0.98
10	0.98	0.98	0.98	0.97	0.98	0.95	0.97	0.97	0.98	1.00	0.97	0.97
11	0.97	0.96	0.98	0.97	0.97	0.97	0.98	0.98	0.97	0.97	1.00	0.97
12	0.98	0.96	0.97	0.98	0.97	0.96	0.98	0.97	0.98	0.97	0.97	1.00

Line 286-291: This is interesting. Is the decrease in RH usually accompanied with changes in other parameters/conditions? Perhaps a change in weather system, cloud cover (i.e. change in radiation), etc? Keep in mind that the features are not always independent variables (certainly not the case in this work, which is perfectly fine). Say, if two features A and B are equally important, the algorithm may give high importance to A (or B) but very low importance to B (or A), thus giving the wrong impression that A (or B) is important but B (or A) is not. I would be a little surprised if the negative response of ozone to RH is really driven by NO_2 uptake under high RH, after all NO_2 is only moderately soluble.

Response: Thank you for your good comments. We agree with you that the features are not always independent variables. RH may change accompanied by other parameters/conditions. We correlated RH with solar radiation (SR) as you suggested. However, RH and SR showed a weak correlation ($r < 0.1$). Meanwhile, we tested the independence of RH and SR as shown in Table R3 or S4, which showed that RH an SR are independent of each other. As shown in Table R2, the algorithm in this work is pretty stable, which indirectly suggests the independence of RH on other parameters, unless great variations of the RI of RH should be observed. In addition, Hu et al. (2011) found that RH was negatively related to the rate constant of HONO formation. Thus, RH also affects O_3 formation by influencing atmospheric OH radicals from HONO photolysis. Therefore, the negative response of O_3 to RH was not just driven by NO_2 uptake under high RH, but also by the deposition of O_3 and the decrease of HONO formation rate constant. We added the discussion in Lines 289-296.

"In addition, it has been shown that RH is negatively related to the rate constant of HONO formation (Hu et al., 2011). Thus, RH might also affect the O_3 formation by influencing atmospheric OH radicals from photolysis of HONO. It should be noted that the negative response of ozone to RH might also be resulted from the dependence of RH on other parameters/conditions, such as SR. However, RH and SR showed a bad correlation ($r < 0.1$). We further tested the dependence of the RI on RH and SR with or without the counterpart as input. The stable RI values (Table S4) mean that RH and SR are independent from each other."

Table R3. Independence test between RH and SR

name	RI value		
	RH and SR as input	RH as input	SR as input
RH	0.68	0.68	/
SR	0.76	/	0.76

Code and data availability: The authors mentioned in the response that the random forest model used in this work is not based on widely used software packages/platforms (such as python/scikit-learn) but is developed in-house in MATLAB (this is impressive, if implemented properly). The vast majority of the dataset used in work is also not publicly accessible as of now. Please refer to the journal policy/guidelines on code and data availability. Given the absolute critical role the random forest model is playing in this work, I strongly recommend that the authors deposit the random forest model code in FAIR (Findable, Accessible, Interoperable, and Reusable)-aligned reliable public repositories.

Response: Thank you. We have deposited the random forest model code on GitHub (https://github.com/z-12/amt-2021-367.git). We have revised it to "The code can be seen in GitHub (https://github.com/z-12/amt-2021-367.git)." in Lines 354-355.

References

Hu, G., Xu, Y., Jia, L. Effects of relative humidity on the characterization of a photochemical smog chamber. J. Environ. Sci., 23, 2013-2018, https://doi.org/10.1016/S1001-0742(10)60665-1, 2011.