Atmospheric Measurement Techniques
Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2021-367-RC1, 2021

Comment on amt-2021-367

Anonymous Referee #1

Referee comment on "Ozone formation sensitivity study using machine learning coupled with the reactivity of volatile organic compound species" by Junlei Zhan et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-367-RC1, 2021

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. However, there are several major issues: (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. (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 Bejing, 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 al. (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 transparaent manner, and to revise the initial VOC calculation, resubmission is recommended. Please see my specific and minor/technical comments below.

Specific concerns:

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

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.

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.

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.

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 themselfs or used that from a certain package (e.g., R, python). If latter, please specify which package is used.

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

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?

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.

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

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.

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.

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

Minor/technical comments:

Line 16: Beijing, China.

Line 21: Define abbreviations upon first appearance. NOx and PM2.5 are not defined. Also x should be subscripted throughout.

Line 22: Delete curves

Line 33: "NO-O3-NO2 cycle" please elaborate. The audience will appreciate clearly described concepts.

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

Line 52-53: "In addition, this method lacks the predictability of O3 concentrations for policy-making." This is a bizarre statement. By definition, observations do not have predictability.

Line 53: OBM can use remote sensing measurements too.

Line 80: has -> have.

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

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.

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?

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

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.

Line 136: Please define “score”.

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

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

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

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 pratically useless. The total VOC levels are less than ~30 ppb (Figure 1F).

Text S2: please define all variables. COH, 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".

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?