Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2022-44-RC2, 2022

Comment on gmd-2022-44

Anonymous Referee #2

Referee comment on "A machine learning methodology for the generation of a parameterization of the hydroxyl radical" by Daniel C. Anderson et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-44-RC2, 2022

The study 'A machine learning methodology for the generation of a parameterization of the hydroxyl radical: ...' by Anderson et al. presents the results of training boosted regression trees on the output of a chemistry-climate model simulation. Specifically, the goal is to predict OH concentrations as a function of meteorological and chemical variables, which could thereafter be used as a parameterization.

Mostly, this is a solid, well-written paper. In particular, most of the technical choices (following the initial choice of algorithm) are well-reasoned. However, I still have a few major and minor technical and scientific concerns, which I would like to see addressed before I would recommend publication. In particular, I am not sure how helpful such a parameterization would be if most of the chemical inputs cannot be simulated interactively. At the very least, this will help to clarify points that other readers will likely find confusing as well.

Major comments:

- My main concern relates to the motivation of building a standalone OH parameterization. I see that the authors cite several studies that used and developed OH parameterizations before and that certain inputs to those parameterizations are themselves simplified (e.g. simply time-varying climatologies). However, given the inputs for the parameterization here (e.g. NO2, ozone, VOCs) would this not pre-determine OH variations hugely, i.e. most of the expected variance in nature would not be included anyway? Is this then really just about capturing OH trends that scale with changes in CO, methane, O3? It seems to me that this is not the variance you primarily learn here, cf. your Gain values which show low contributions from e.g. CO, or CH4. If one only cares about the effect of OH on CH4, would a simpler way not be to predict long-term CH4 changes as a function of the anyway prescribed changes in CO and O3?
- The title is longer than necessary. Maybe consider shortening it? For example, are both

'methodology' and 'tool' needed in the same title? I am also not sure the main application of this tool would actually be in chemistry-climate model simulations. Surely, if you already run an interactive chemistry scheme it would not make a big difference to just replace OH by a parameterization? Maybe you mean something different, but it could confuse potential readers.

- Training and evaluating a parameterization offline is a completely different animal from predicting 'online' ie in operational mode, especially if suddenly the inputs won't be provided by a consistent interactive chemistry model anymore. By now this is a well-known fact about machine learning parameterizations. This must be highlighted somewhere, i.e. that the study here is an offline test of the principle that requires further validation before being used operationally, especially if the OH feedbacks onto the system itself somehow (which I guess it ultimately should according to your paper title).

Minor comments:

- l.94-98: Please also cite

Nowack et al. Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. Environmental Research Letters 13, 104016 (2018). https://iopscience.iop.org/article/10.1088/1748-9326/aae2be

Since this was the first study to suggest machine learning parameterizations in atmospheric chemistry. Note that the authors found that ridge regression outperformed random forest regression in their case, which might have to do with ridge regression allowing for some degree of extrapolation, which you find is a potential issue here (e.g. see also Nowack et al. AMT 2021 https://amt.copernicus.org/articles/14/5637/2021/amt-14-5637-2021.pdf).

Only trying out one algorithm is actually one of the main shortcomings of this study, given the aim to publish in GMD. Why not try at least a few different methods to compare their performance? This should be computationally feasible and it is well known that there is 'no free lunch' in machine learning, so just arguing that a method is chosen because others did the same before, often in very different applications, is certainly the (too) easy way out.

Another random forest application in atmospheric chemistry worth mentioning is Sherwen et al. ESSD (2019): https://essd.copernicus.org/articles/11/1239/2019/

- l. 101-104: yes, although the SHAP values could also be used for other methods, of

course. Linear machine learning models such as ridge and Lasso are of course even more interpretable.

- l. 106-113: my main concern here is that you train your parameterization on a single simulation and mention, correctly, that the parameterization cannot extrapolate outside the training domain. Given that we are usually interested in transient climate change scenarios, I would strongly recommend training the parameterization on a wide range of simulations (ideally at least one extremely strong forcing and one strong mitigation scenario). This would mean that re-training is expensive, unless the data already exists. Maybe that should be highlighted instead, i.e. that you could usually learn from existing simulations that will be run anyway?
- l. 122: that's a key concern: I doubt your current parameterization would work under climate change conditions, which makes it less useful. I think this point should be made very clear and highlighted prominently in the paper, ie that this is just a recipe to learn a parameterization but not a read-to-go product. Not all readers will be aware of this issue.
- l. 152: why not 2019? To avoid maximum extrapolation? I think it is good that you left out five years from the training data completely, this will reduce the risk of inflating performance measures due to spatial or temporal autocorrelation. I assume I understand your motivation correctly?
- Table 1: at this point it is still unclear to me how you consider the vertical dimension, so this should be clarified earlier on (unless I overread it somehow). Do you predict OH in each tropospheric grid box (from the near-surface to the upper troposphere)? Does you include only ozone, CO etc values from the same grid box you are trying to predict, or would, for monthly-mean training data, it not be best to include at least surrounding spatial predictors? I am also not sure if, especially for surface OH, it makes sense to build one parameterization that predicts all grid points simultaneously? Wouldn't there be a spatial dependency of how e.g. ozone and OH relate, or do you think this will be sufficiently conditioned out by the other variables you include?
- l. 170: why separate by month and not by location? Should not June at a SH grid location be more different from June in the NH than, say, from July at the same grid point in the SH?
- l. 224: Random forests are known to often overpredict small values. Still wondering if this might partly be a result of throwing all data points independent of locations in one training basket?
- Figure 1: is this now tropospheric column-averaged OH? Surface OH? Please clarify. I read on: I now see that you say in the text that his is the PBL-500hPa average. Still, would add that info to the figure caption, too. Any particular reason to avoid the boundary layer? Is the spatial invariance not a good assumption there?
- At this point: do I understand correctly that you train and predict monthly-mean data? Your description in the data section was eventually somewhat confusing. So, do you train on monthly-mean data but use those functions to also predict daily OH? Might be worth highlighting again in the figure caption, for clarity.
- l .250-253: implies that you tried both, ie training on daily and monthly data. Here you say that daily performs better, but above you seem to imply that daily suffers from other problems with extrapolation. What would you recommend then? Please clarify.
- Figure 2: clarify that this is based on daily predictions (ie based on the model trained on daily data).
- Figure 3 should be larger, otherwise this becomes hard to read.
- Figure 5 and l. 351: again is this based on the model trained on daily or monthly-mean data? In any case, I would assume that CH4 has low predictability as it shows low variability on relatively short daily and monthly timescales, where other factors driving internal variability (rather than trends) are providing greater contributions to the predictions, thus showing greater feature importances. I suppose that would explain your initially maybe somewhat surprising result. If you predicted longer-term averages (e.g. decadal) I guess the picture would change dramatically. Anyway, this links to my major comment above and if much of the variability you capture will be negated if

chemical inputs are derived from much simpler climatologies.

- l. 447: I see – was the motivation to exclude 2016 the specific El Nino event?
- l. 542-583: I think this is a very important discussion. Random forests cannot extrapolate, so this behaviour is not surprising (e.g. Nowack et al. AMT 2021). As a more general point, I would rephrase the entire paper, also the abstract, in the sense that you present a way to learn relationships, i.e. to show that this is possible, rather than a ready-to-go parameterization, see also my other points above.
- Section 4.1.2: again, not surprising because random forests cannot extrapolate. So, if you change the distribution mean, shape, and ranges, you will be in trouble. Again, selling your results as a recipe rather than a ready product, would address this point immediately.