Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-772-RC1, 2019

Atmospheric Chemistry and Physics Discussions

Interactive comment on "A Machine Learning Examination of Hydroxyl Radical Differences Among Model Simulations for CCMI-1" *by* Julie M. Nicely et al.

Peer Johannes Nowack (Referee)

p.nowack@imperial.ac.uk

Received and published: 2 October 2019

The paper by Nicely et al. uses a neural network approach to infer drivers of differences in OH/methane liftetimes among chemistry-climate models. In addition, the approach is used to understand modelled historic trends and variability in these variables. The method itself has been applied in similar form before (cf. Nicely 2017), but here it is applied to a novel set of specified dynamics CCMI simulations.

Overall this paper is a nice example of how machine learning can be used to provide novel insights into chemistry-climate model differences and I enjoyed reading it. I would therefore definitely recommend rapid publication subject some revisions and

clarifications concerning my comments listed below.

Major comments:

- The use of neural nets and especially their cross-validation requires further motivation and explanation. I know this can feel like unnecessary repetition to the authors given that the method has been described previously, but it is an essential aspect due to the central role of the method here. For example, when I first read the paper I was entirely unclear if all results might be subject to overfitting and if the sampling was done in space or time as well as how the data was split into training, cross-validation and test datasets; an essential aspect of any machine learning application. I now understand from reading the other paper that probably regressions were fit on an 80%/10%/10% split of the year 2000, using each grid cell as one sample for a month (rather than samples being ordered by time). Is this still valid? Is early-stopping really the only method you used to manage the bias-variance trade-off? This point is particularly important as evaluation results are given only for the year 2000, which as mentioned is used for training. Given that the year was used for training it would not be surprising if the neural net can fit the data almost arbitrarily well if overfitting wasn't sufficiently counteracted. Maybe show results/evaluate for all years that you did not use for training? I would also explicitly mention the sample size for each dataset (all models are interpolated to the same resolution?).

- I would like an additional explanation of why neural nets were used in the first place. I know they can model complex non-linear functions (which is one point that could be mentioned), but there are many algorithms that can do the same but would probably be more suited for inference tasks such as the one attempted here. Random forests, for example, would immediately provide feature importances for the regression models themselves and it would be easier to test dependencies between correlated variables (e.g. ozone, T, humidity) where it is

unclear what is cause and effect. I do not ask for a refit with different algorithms, but it could be mentioned in terms of future work/context.

- some more reflection on the role of the nudged dynamics: the authors mention that one of the reasons why temperature is less important in explaining inter-model differences is the fixation to a common atmospheric background state by nudging. Alternatively, correlations with other variables such as ozone are offered as an explanation. Could the same not be said about water vapour? Maybe this would also explain why it is suddenly so much more important (relatively) to explain variability? What did you observe in this respect for the free-running simulations?

- the randomness of neural networks: it seems that only one network is fit per model. Unfortunately, neural networks behave somewhat randomly, which is essentially the result of many different local minima in the cost function that can be found during the weight optimization process. Therefore, I would expect that the networks for each model would already be different due to different random initializations of the networks even if the chemistry models would be identical. I would strongly encourage the authors to test the relative importance of this randomness aspect compared to the actual inter-model differences. For example, they could train five-ten neural networks for two of the models (subject to an objective optimization/early stopping procedure) and show the spread in the results when these different network realizations of the two models are compared (instead of only one realization for each). No need to get started with different network architectures, which would similarly affect the results, I assume.

Minor comments:

- p. 4, l.107-109: revise second part of the sentence.

- p. 4/5; model simulations: since UV fluxes and stratospheric ozone are discussed maybe briefly mention if all/which models include interactive stratospheric chemistry, or how it is treated otherwise.

- section 3.1 I think there should be more detail here; essentially another small subsection on the cross-validation method.

- p. 5 l. 161: 'mutually exclusive' - what do you mean by that here?

- l. 165-170: Maybe try a variation of the input features? The cross-correlations are indeed an obvious problem for the interpretation. Did you consider fitting two different networks, e.g. one with JO1D, one with column ozone and consider how well they do on the cross-validation dataset? I am also wondering how these different networks would perform in different atmospheric regimes, e.g. column ozone being more important in the upper troposphere. JO1D (including clouds) becoming relatively more important in the lower troposphere? Can a single network for all grid cells capture these different regimes appropriately?

- l.228: performance for the year 2000 is strong – but this is the training year. Should the goal not be to evaluate on out-of-sample years. Maybe show an error plot for all years? I assume it gets worse the further one moves away from the training year, partly due to the extrapolation error?

- general remark on the extrapolation issues: could you give an estimate of how often you had to correct values in this way for each comparison/model (e.g. percentage o cases depending on the year)? This would give the reader a better impression of how important this factor is when considering the results. In addition, did you ever test how linear/non-linear the regression relationships really are? Maybe linear regression algorithms such as Lasso/Ridge would actually circumvent all these issues by being able to extrapolate better and still extract

feature importances in a sophisticated enough manner (the resulting regressions would also be easier to interpret).

- l. 484: maybe I approach this one too naively, but why would I expect to model a CH4 trend if CH4 is normalized by its maximum value in each year? I assume the maximum value shows a trend somewhat proportional to the average trend?

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-772, 2019.