Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-972-RC1, 2020

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Developing a novel hybrid model for the estimation of surface 8-h ozone (O_3) across the remote Tibetan Plateau during 2005–2018" *by* Rui Li et al.

Anonymous Referee #2

Received and published: 25 January 2020

This paper attempts to create a machine-learned statistical model to estimate 8-h surface ozone concentrations where few measurements exist (Tibetan plateau). The model formulation (RF-GAM) is interesting and the subject fits within the scope of ACP. There are several flaws with the presentation of this work, namely the lack of significance tests and alternate performance metrics in addition to English grammatical errors throughout. However, I am optimistic the authors can correct these issues and I believe this work will be a very important addition to the scientific literature of air pollution in rural regions. 1. Does the paper address relevant scientific questions within the scope of ACP? This paper attempts to create a machine-learned statistical model to estimate 8-h surface ozone concentrations for the Tibetan plateau. The RF-GAM

model is interesting and the subject fits within the scope of ACP.

2. Does the paper present novel concepts, ideas, tools, or data? The machine learning model formulation is interesting and the application of the GAM to remove autocorrelation in the time series data is novel.

3. Are substantial conclusions reached?

Not particularly, the performance metrics presented in the work are not convincing. The authors present RMSE as an absolute measure of performance yet the best performing model had an RMSE of 14.41 μg/m3, which is greater than 10% error relative to the WHOI 8-h ozone critical value (100 μg/m3) they are using to determine nonattainment. There is very little discussion about significant differences between performance due to model architecture and between seasons and years. Greater discussion of model uncertainty needs to be had.

This study addresses an important topic that lacks much attention in scientific literature, but the performance of the model is still considerably lacking.

4. Are the scientific methods and assumptions valid and clearly outlined? Scientific methods and data preprocessing are outlined well. Figure 2 presents a workflow method for the machine learning model, though including an actual architectural schematic of the final model would be better (e.g. what does the random forest look like in terms of the number of trees, nodes, etc.). Furthermore, none of the other machine learning models were given an architectural model description, i.e., how many nodes in the neural network, activation functions, etc. Machine learning is a burgeoning field with many nascent applications. Therefore, this paper would benefit going into greater detail about what other machine learning methods they tried in detail so that others could learn from their attempts.

5. Are the results sufficient to support the interpretations and conclusions? (cite sections)

There must be a greater statistical argument presented to support the interpretations of the authors. Oftentimes the authors present only an R2 metric or an RMSE value to evaluate the goodness of fit among conditions. These performance measures are relatively obfuscating as even though ozone might increase or decrease between temporal/spatial extents, these changes, I suspect, are statistically insignificant. There are practically no significance tests conducted. These performance metrics must be put into greater context. Perhaps including confidence intervals (rather than only standard deviations) would also help.

For the variable importance section 3.2, the authors should present an error covariance matrix for the inputs. For random forest algorithms, the variable importance determined by the model is meaningless if the variables are co-linear.

Furthermore, oftentimes the explanations for the concentration of ozone are too speculative (e.g. lines 334-351). There are numerous mentions of NOx and VOCs contributing to ozone loss or formation, yet this analysis cannot address such concerns directly as the model does not take into account these variables. These explanations are well thought out but exceed the scope of what this model can answer. Perhaps place them in the discussion rather than with the results.

In addition to RMSE, an absolute metric, perhaps the authors should consider reporting a relative metric such as percent error in tandem. R2 values of 0.60 and RMSE values of 14 μg/m3 are not readily obvious/interpretable if this is adequate performance. Also, conducting significance tests between different models might be beneficial. I am not convinced that RF and XGBoost are markedly different. The addition of the GAM seems like a different model architecture than typical machine learning models. That is, why not attach the GAM to the XGBoost and see how that performs? It is fine if the authors wanted to create an RF-GAM model from the outset, but the comparison between the other models does not seem very robust.

6. Is the overall presentation well structured and clear? The overall structure and flow

of the paper are coherent, though oftentimes the English is not clear.

7. Is the language fluent and precise? This study needs considerable corrections to numerous grammatical and language errors. There are far too often awkward phrasings and mixed tenses.

8. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?

Many of the figures are too small to view. Needs to be increased in size.

9. Are the number and quality of references appropriate? Great overview and command of the literature.

Following are some line comments, though not all-inclusive: 44: a few times you use the word 'beneficial' to describe conditions that create pollution, perhaps consider different word choices as this is a bit awkward. Also found on line 127.

102-103: "decision tree models such as random 101forest (RF) and extreme gradient boosting (XGBoost) strike a perfect balance between 102prediction performance and computing cost". This seems to be a wide sweeping claim without a reference. Decision trees may be faster to train than neural networks, but they are also slower to evaluate. That is, decision trees are slow to train and slow to test, whereas neural networks are very slow to train and very fast to test. Consider revising your claim.

135-138: awkward, confusing phrasing

190: Still not sure why you included GDP as an input variable. You write that you integrated it because "these data were available each five years". Please provide a scientific explanation for the inclusion of this variable.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-972, 2020.