Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC1
https://doi.org/10.5194/acp-2021-634-RC1, 2021


# Comment on acp-2021-634

Anonymous Referee #2

Referee comment on "Interpreting machine learning prediction of fire emissions and comparison with FireMIP process-based models" by Sally S.-C. Wang et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-634-RC1, 2021

1. Interpretable ML or interpreting ML

There is a major difference in internally interpretable ML models versus trying to interpret ML outcomes with analytic tools [Rudin 2019]. I think this study belongs to the second category.

In that case, the title is really confusing, first XGBoost model is explainable that one can easily derive variable importance. However, SHAP was finally used to interpret XBGoost.

Interpreting an interpretable ML model (XGBoost) with an interpreting tool (SHAP) is a little bit weird. Why not just used XGBoost derived variable importance? What's the value of SHAP? If SHAP is really necessary and could give us a better understanding of XGBoost, then at least the title needs to be updated to: e.g., "Interpreting ML prediction of fire emission xxx". And focus more on why SHAP is better than XGBoost internal variable ranking.

2. What can we learn and to inform future development

As highlighted in the abstract that one of the objectives was to inform model future development. However, it wasn't sufficiently discussed and there is no clear conclusion on e.g., which part of the process-based model needs major development? What is missed in process-based models?

3. Different time scales: long-term trend, inter-annual variability, sub-seasonal dynamics

For different time scale, one would expect different dominant driver for wildfire burn and PM2.5 emissions. For example at the sub-seasonal scale, climate may play a more important role, while the long-term trends may be more affected by human activities. I wonder is it possible to carry out some experiments to better interpret the ML outputs across different scales? For example, detrend the time series (long-term trend) and use ML to predict interannual variability and compare the variable importance with the original ML models?

4. Uncertainties in data and ML model training/prediction

There are multiple existing datasets (e.g., Fire Atlas, Fire CCI). One potential issue of training/validating an ML model using only GFED data is that the ML predictions are subject to GFED uncertainties. If possible, a comparison of GFED with other products and even better, applying GFED emission factors to other BA products, then one can train/validate ML model towards multiple datasets of fire emissions, include the data uncertainties in the cost function.

Others:

L23: xxx, which may be explained by the coarse spatial resolutions of the processed-based models or atmospheric forcing data or limitations in model parameterizations for capturing the effects of Santa Ana winds on fire activity. This statement is not helpful. What is the real reason why the FireMIP model did not capture bimodal peak emissions? For example, one can check wind fields in GSWP3 forcings or CRUNCEP forcings to verify the existence of Santa Ana winds.

L108: How was emission factor data estimated? Are the emission factors PFT dependent or constant across the whole US?

Section 2.2. What are the differences in input variables used by FireMIP and ML model?

For a fair comparison, it will be good to make sure ML and FireMIP models used the same input variables. But, if not, what are the implications of using different input variables, and how do they contribute to the ML and FireMIP model differences?

L171: LAI is a poor indicator of biomass since the majority of the biomass comes from the stem. As long as the canopy is closed, the growth of vegetation biomass no longer link to LAI.

L173: Worth first evaluate CLM fuel load with existing present-day biomass datasets, then apply it to ML model.

L202: FireMIP models used cru-ncep, while ML model used NARR and gridMET, ML's fuel load input was simulated with GSWP3 forcings? The differences in climate forcings make the comparison less valuable, especially when forcing uncertainties dominated the comparison. Maybe one can eliminate the forcing uncertainties by first surrogate FireMIP with ML models and replace CRUNCEP forcing with the GSWP3 forcings.

Section 2.3. Has the FireMIP model sufficiently tuned using GFED data? Since the ML model was maximally tuned towards the GFED dataset, it's important to clarify whether or not FireMIP also tuned towards GFED. Otherwise, it's expected that the ML model would outperform FireMIP models.

Figure 4,5. The results will more meaningful if the regression was done for only the peak fire months (or fire season), given that emissions only happened during fire season.

Reference

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.