Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-261-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Quantitative assessment of fire and vegetation properties in historical simulations with fire-enabled vegetation models from the Fire Model Intercomparison Project" *by* Stijn Hantson et al.

Anonymous Referee #1

Received and published: 5 February 2020

The authors provide a detailed glimpse into the successes and struggles of global fire modeling efforts, and quantitatively try to isolate the most pressing challenges for both individual fire models and the fire modeling community as a whole by using a benchmarking method of comparisons with observations. Particularly interesting is that the authors highlight how sensitive the benchmarking results are to how vegetation (fuel load) is captured or simulated in any particular fire model. I think the paper should be published after a few minor revisions and/or author responses/clarifications to my concerns below.

Comments

Title: this is a confusing title because nowhere in the paper are the historical FireMIP simulation results discussed. Line 145 and essentially all the figures point out that only present day results are analysed. "Historical" in the CMIP framework usually refers to simulation periods that extend from about 1850 to present day. I would strongly suggest changing the title to better capture the scope of the analysis the authors undertook.

Lines 95-104: this paragraph is difficult to follow relative to the analysis of FireMIP output. Are the authors trying to say that benchmarking allows for a more systematic evaluation of models so that a hierarchy can be quantifiably justified? If so, I suggest that the authors add or clarify this in the text to make it clear to readers that this is why the authors chose to raise this discussion point. Alternatively, the authors could shorten or delete the paragraph altogether, because while they raise the point of ending model democracy, it stands in contrast with the conclusions of the study, where the authors say the "no model clearly outperforms all other models" which seems to be avoiding the issue of hierarchical treatment of the fire models. If this group of authors cannot ascribe a hierarchy to global fire models, then I think they miss the chance to advance the conversation from the perspective of their collective expertise. By this, I mean that I, as the reader, can walk away from the paper with useful benchmarks and metrics, but that I will also then evaluate model quality on my own because the authors did not. My conclusion is that while the benchmarks are great to have, the results in Figure 2 and Table 2 clearly show that GlobFIRM and MC2 output should not be considered equally alongside output from other models.

Paragraph at line 166: Certainly there are observational uncertainties, but the Global Fire Atlas and other studies about fire products (GFED papers and MODIS papers, at least) have made a solid effort to quantify uncertainties – what do the authors suggest is enough in terms of validation of the observations? Some specific problems I have with the paragraph: In line 169, saying "large uncertainties still remain for most variables" is too vague. Which variables? How large, or large compared with what? To me,

it seems that fire models have larger uncertainty than the observations. I would argue that the results in this paper suggest that model uncertainty does not arise from a lack of observations, but rather, the model uncertainty is largely due to poor simulations of biomass. While this paragraph makes it sound like models are waiting for observations of bulk properties, it is more accurate to say that the fire models do not have the fuel process simulated correctly. These are two different issues that should not be about a lack of observational constraints. I suggest the paragraph be shortened a sentence or two so that the focus of the paper remains on evaluation of model output, and not observations. The authors could simply point out that burnt area, biomass, and fire emissions estimates vary and uncertainty is still being characterized, and cite appropriate papers. To me, this paper is about the benchmarking results, and the fact that observations have weaknesses too should be relegated to a side note with citations.

Table 3: Why are the benchmarking scores for the Mean null model often equal to 1? Is this an artifact of the calculation itself? If so, wouldn't this detract from the utility of using the Mean null model as a point of comparison with fire model benchmark scores for those fire variables?

Paragraph at line 229: The text discussion seems inconsistent with the results in Table 3. I may be misunderstanding the reason for the benchmarking scores for the Mean and Random null models, but my interpretation is that those Mean and Random null model benchmarking scores are the target to beat. If a fire model beats that benchmark score, then my interpretation is that that particular fire model performs better than the null model. Is that a correct interpretation? If so, then there seem to be some inconsistencies between the text and Table 3 as follows.

Paragraph at line 229: Specifically, one sentence states "The models capture the timing of the peak fire season reasonably well, with all of the models performing better than both null models for seasonal phase in burnt area" but many of the fire models have benchmark scores greater than the Random null model, so why do the authors say "all"?

Paragraph at line 229: Another sentence states "all of the FireMIP models perform worse than both null models for seasonal concentration of burnt area, independent of the reference burnt area dataset" but looking at Table 3, almost all of the fire model benchmark scores are less than the benchmark scores for the Random null model, with the exception being JULES-INFERNO vs FireCCI40. Wouldn't this mean that the comparisons are all better than the Random null model?

Future model development section: I would suggest that the authors propose mechanisms that fire models should include (crops, prescribed biogeography), and reflect on both why some fire models do not include those mechanisms already, and whether the future of fire model development will include those mechanisms. Or perhaps this is discussed in other FireMIP papers already? Also, the authors might provide a broader perspective in this section by discussing whether there are global fire models currently in use that did not participate in FireMIP but do include features that the benchmarking results in this study highlight as particularly weak. For example, Pfeiffer et al's LPJ-LMFire model https://www.geosci-model-dev.net/6/643/2013/ includes representation of human use of fire in a novel way.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-261, 2020.