Author response to Anonymous Referee #1 review

The authors thank referee #1 for this considered comments and constructive suggestions.

Below we provide a detailed response in italic to each comment.

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.

The simulations we examined are indeed historical, in the sense that they were run from 1700 CE to the present day, although we only evaluate them in the recent past because of the availability of data. But we agree the title might imply evaluation over a longer period, and we will change it to: “Quantitative assessment of fire and vegetation properties in simulations with fire-enabled vegetation models from the Fire Model Intercomparison Project”.

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.

We agree that the GlobFIRM and MC2 simulations are poor and not comparable to the other simulations in the FireMIP ensemble, and indeed we state this (lines 369-370). The reviewer indeed interprets the objective of this paragraph correctly as we think that establishing a hierarchy of model's ability to simulate fire is important (and hence we would like to keep the paragraph explaining that this is one of the goals of benchmarking) and we should have made a stronger statement about this in the abstract and conclusion. We will modify the text as follows:

- *Line 57 et seq. "The two older fire models included in the FireMIP ensemble (LPJ-GUESS-GlobFIRM, MC2) clearly perform less well globally than other models, but it is difficult to distinguish between the remaining ensemble members: some of these models are better at representing certain aspects of the fire regime, none clearly outperforms all other models across the full range of variables assessed."*
- *Line 319: "Our evaluation suggests that LPJ-GUESS-GlobFIRM and MC2 produce substantially poorer simulations of burnt area and its inter-annual variability than other models in the FireMIP ensemble. These are both older models, developed before the availability of global burnt area products (in the case of LPJ-GUESS-GlobFIRM) or calibrated regionally and not designed to run at global scale (MC2). While the other models perform better in simulating fire properties, there is no single model that outperforms other models across the full range of fire and vegetation benchmarks examined here. Model structure does not explain the differences in model performance."*
- *We furthermore included an extra paragraph at the end of the discussion to cover this point: "Our analysis demonstrates that benchmarking scores provide an objective measure of model performance and can be used to identify models that might negatively impact on a multi-model mean and so exclude these from further analysis (e.g. LPJ-GUESS-GlobFIRM, MC2). At the moment, a further ranking is more difficult because no model clearly outperforms all other models. Still, some FireMIP models are better at representing some aspects of the fire regime compared to others. Hence, when using FireMIP output for future analyses, one could weigh the different models based on the score for the variable of interest, thus giving more weight to models which perform better for these variables."*

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.

We agree that the focus of this paper should be on the evaluation of the model results. Our intention in this paragraph was definitely not to critique the groups producing different fire datasets or to imply that they are not trying to provide both theoretical (Brennan et al., 2019) and practical uncertainty estimates (e.g. Giglio et al., 2013), but to explain why we do not take account of observational uncertainties in our comparisons. *We agree with the reviewer that the uncertainty in model output exceeds the uncertainty of existing datasets and we agree that this paragraph might distract, and we will shorten it drastically and reduced it to its essence, rewriting it as follows:*

"Ideally, model benchmarking should take account of uncertainties in the observations, for example by down-weighting less reliable data sets (e.g. Collier et al. 2018). However, observational uncertainties are not reported for some of the data sets used here (e.g. vegetation carbon). Furthermore, some of the data sets (e.g. emissions) involve modelled relationships; there has been little formal assessment of the choice of model on the resultant observational uncertainty. While we use multiple datasets when available (e.g. for burnt area, where there are large differences between the products), in an attempt to integrate observational uncertainty in our evaluations, it seems premature to incorporate uncertainty in the benchmark data sets in a formal sense when calculating the benchmarking scores."

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?

The normalized mean error (NME) is constructed in such a way as to normalize the scores against an objective background so that the mean null model results in a score = 1 (Kelley et al., 2013, Biogeosciences 10: 3313-3340)*. This is not an artifact but a design feature of the metric to make the interpretation of the results more intuitive compared to other error metrics. All the values shown as less than 1 in Table 3 are for seasonal phase and are calculated using the mean Phase Difference metric, which is not constrained in the same way. Since our description of the metrics is not clear, and also in response to comments by the second reviewer, we have rewritten the section of text describing the metrics and the null models as follows:*

"To assess model ability to reproduce spatial patterns in a variable, we use the normalised mean error (NME):

$$NME = \frac{\sum A_i |obs_i - sim_i|}{\sum A_i |obs_i - \overline{obs}|} \qquad (1)$$

where the difference between observations (obs) and simulation (sim) are summed over all cells (i) weighted by cell area (A_i) and normalized by the average distance from the mean of the observations ($\overline{obs}$). Since NME is proportional to mean absolute errors, the smaller the NME value the better the model performance. A score of 0 represents a perfect match to observations. NME has no upper bound.

NME can be sensitive to the simulated magnitude of the variable. To take this into account in comparisons, we removed the influence of biases in the mean and variance between model results and each reference dataset. This has the further desirable property of limiting the impact of observational uncertainties in the reference datasets on the comparisons. Although we focus on benchmarking results after removing biases in the mean and variance, the scores for comparisons before this procedure (and for comparisons after removing mean biases only) are given in Supplementary Information S2.

To assess model ability to reproduce seasonal patterns in a variable, we focused on seasonal concentration (roughly equivalent to the inverse of season length) and seasonal phase (or timing). We calculated a mean seasonal "vector" for each observed and simulated location based on the monthly distribution of the variable through the year. The concentration is the length of this vector compared to the annual value, and ranges between 0 when the variable is distributed evenly throughout the year and 1 when the season is confined to a single month. The phase is indicated by the direction of the vector. Observed and modelled concentrations were compared using NME. Phase is compared using the Mean Phase Difference (MPD) metric (see Supplementary Information S2). Again, for NME, a score of 0 represents a perfect match to observations and

there is no upper bound. MPD has a maximum value of 1 when all cells have a maximum phase mismatch of 6 months. Seasonality metrics could not be calculated for three models (LPJ-GUESS-GlobFIRM, LPJ-GUESS-SIMFIRE-BLAZE, MC2), either because they do not simulate the seasonal cycle or because they did not provide these outputs. We did not use FireCC4.0 to assess seasonality or interannual variability (IAV) in burnt area because it has a much shorter times series than the other burnt area products.

Model scores are interpreted by comparing them to two null models (Kelley et al., 2013). The "mean" null model compares each benchmark dataset to a dataset of the same size created using the mean value of all the observations. The mean null model for NME always has a value of 1 because the metric is normalised by the mean difference. The mean null model for MPD is based on the mean direction across all observations, and therefore the value can vary and is always less than 1. The "randomly-resampled" null model compares the benchmark data set to these observations resampled 1000 times without replacement (Table 3). The "randomly-resampled" null model is normally worse than the mean null model for NME comparisons. For MPD, the mean will be better than the random null model when most grid cells show the same phase. A detailed description of the benchmarking metrics is given in the Supplementary Information S2. "

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.

Your interpretation is correct. We stated this in the original manuscript (lines 192-198) but we have now rewritten the section on the metrics and their interpretation (as described above) and hope this is now clearer. The confusions between the Table and the text are due to mistakes on our part in the description and we have now corrected these, as explained below.

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"?

This sentence should have read "The models capture the timing of the peak fire season reasonably well, with all of the models performing better than the mean null model for seasonal phase in burnt area". And have added additionally: "The models also frequently perform better than the random null model, with all models performing better against GFED4.".

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?

This should have read "mean null model" instead of "both null models" and has been corrected.

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.

We agree that the title of this section is somewhat misleading, since we do not believe it is possible, as yet, to prescribe exactly the steps that would yield an improved fire model. Our intention here was to point to areas which need to be investigated further because the benchmarking identifies them as weaknesses in the current models. It would be possible, for example, to include crops into the models or human use of fire (as in LPJ-LMFire). However, the current parameterizations of agricultural fires are relatively simple and generally not based on rigorous data analysis. And indeed, as the ongoing discussion about the impacts of anthropogenic activity on fire trends shows, our understanding of human-fire interactions is very incomplete. Similarly, we have identified the ability to reproduce vegetation properties and hence fuel loads as an area where the models do not perform well -- but again, this is an active area of research and, as yet, there is no agreed way forward. However, we agree that it would be valuable to point out which processes are already implemented in different fire models not participating in FireMIP (including LPJ-LMFire) and will adapt the discussion section in different points accordingly. We will also re-title this section simply as: Discussion.

References

Brennan, J., Gómez-Dans, J. L., Disney, M., and Lewis, P.: Theoretical uncertainties for global satellite-derived burned area estimates, Biogeosciences, 16, 3147-3164, 10.5194/bg-16-3147-2019, 2019.

Hall, J. V., Loboda, T. V., Giglio, L., and McCarty, G. W.: A MODIS-based burned area assessment for Russian croplands: Mapping requirements and challenges, Remote sensing of environment, 184, 506-521, http://dx.doi.org/10.1016/j.rse.2016.07.022, 2016.

Roteta, E., Bastarrika, A., Padilla, M., Storm, T., and Chuvieco, E.: Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa, Remote Sensing of Environment, 222, 1-17, https://doi.org/10.1016/j.rse.2018.12.011, 2019.