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.

Author response to Sergey Venevsky's review

The authors thank Sergey Venevsky *for his constructive comments. Below we provide a detailed respond in italics to each comment and suggestion.*

The paper presents evaluation of historical simulations made for the nine FireMIP models with regard to fire and vegetation properties. This is very important and necessary inter comparison study aimed to support and move further global fire modelling activities. Both methods and results are clearly presented and scientifically sounded and proven. The only methodological weakness is a bit short period of comparison of the models with the observations which could be clearly longer. I think that Discussion in this paper is the weakest part and needs more effort to make conclusions from the model intercomparison to be stronger and more clear. In particularly 1. The part about relation of areas burnt to vegetation production (lines 387 -396 and 405-411) should describe in more details why and how exactly fuel load influence burnt areas in the FIREmip nine models 2. Part on influence of areas burnt upon GPP/NPP is absent and should be added.

The global fire models participating in FireMIP use output from the vegetation models as input (e.g. different fuel loads, fAPAR, etc.) to a range of fire processes, depending on the fire model parameterization and structure. A detailed description of the fire model structure, and he interactions with the vegetation model is outside the scope of this paper and has been described in Hantson et al. (2016) and Rabin et al. (2017). However, we acknowledge the need to indicate that this interaction is present and will now modify the text to make this clearer (line 387):

"Vegetation type and stocks are input variables for the fire models, influencing fire ignition and spread in the process-based models and determining simulated burnt area in the empirical models. The occurrence of fire can, in turn, affect the vegetation type, simulated vegetation productivity (i.e. GPP, NPP) and hence the amount and seasonality of fuel build up. Our results indicate that inter-model differences in burnt area are related to differences in simulated vegetation productivity and carbon stocks. Seasonal fuel build-up and senescence is an important driver of global burnt area. Furthermore, we find that models which are better at representing the seasonality of vegetation production are also better at representing the spatial pattern in burnt area."

Comments/questions/suggestions:

Line 66 – "Willdfires and anthropogenic fires" – how you define and classify these types of fires in global models? Further on no clarifications for this important question. . .

We make the distinction here between lightning-ignited fires and fires set by humans. Some models include human ignitions as a function of population density; one model (CLM) also includes cropland fires. The specific parameterizations included in the FireMIP models are discussed in the FireMIP protocol papers (Hantson et al., 2016; Rabin et al., 2017). Since we are not documenting the models here, but only evaluating their performance we do not think that a full discussion of these parameterizations is warranted. However, we will modify this sentence to clarify this point:

"However, the representation of both lightning-ignited fires and anthropogenic fires (including cropland fires) varies greatly in global fire models."

Line 126 Lightning data 1900-1920.population density and land use 1700 where do these data come from?

The source of the data sets is given in the FireMIP protocol, which we cite here (line 133). However, we will modify this to make it clear that the information about all the drivers of the simulation are available there, as follows: "(see Rabin et al., 2017 for description of the modelling protocol and the sources of the input data for the experiments).".

Line 132 The baseline FireMIP simulation is a transient experiment starting in 1700 CE and continuing to 2013. Why simulation is only up to 2013 and comparison is only for 2002-2012? Can simulation be somehow extended to include recent years? Similarly, inter-comparison only for decade looks not so sounded, for example 1997/1998 El Ninio years are out. . . Which climate data was used?

The simulation protocol was drawn up in 2016 and at that stage there was no readily available driving data post-2013. The choice of the interval 2002-2012 for benchmarking was motivated by two considerations. First, 2002 was the first year when both MODIS sensors were operative and hence a temporally coherent high quality global burnt area dataset is available. Before 2002, and especially before 2000, the quality of the GFED4 archive is much lower. Second, two modelling groups did not run the final year of the simulation (i.e. 2013). We therefore excluded this year in order to keep the evaluation between the models as consistent as possible. CRU-NCEP climate was used as forcing data for the simulations (see Rabin et al., 2017).

We agree that there are several other evaluations of fire models that could be made and certainly testing how well they emulate the response to El Nino variability would be an interesting case study. However, this is out of scope for the current round of FireMIP.

Lines 150-159 What was the principle of selection of all these datasets? Why no global water cycle related datasets (e.g. runoff) where selected? Water status is obviously important for both fire and vegetation, I would compare at least also runoff for 2002-2012

We agree that there are many other opportunities for evaluating fire models, and it would be possible to include data sets such as runoff. However, there are a number of issues that led us not to include runoff in our evaluation: (1) runoff data in the most fire-prone regions (e.g. the Sahel) are not very reliable; (2) gauged runoff is integrated property, and thus requires models to incorporate some form of hydrological routing scheme; and (3) it is not the most relevant hydrological variable for fire -- soil moisture or litter moisture would be useful but are very heterogenous and thus global evaluation would be difficult. We agree that we could usefully modify this sentence to explain our current focus and will modify the text to read:
"Model performance was evaluated using site-based and remotely sensed global data sets of fire occurrence, fire-related emissions and vegetation properties (Figure 1; Figure S1). We include vegetation variables (e.g. GPP, NPP, biomass, LAI) because previous analyses have indicated that they are critical for simulating fire occurrence and behaviour (Forkel et al., 2019a; Teckentrup et al.,2019) and there are global data sets available. We did not consider parameters such as soil or litter moisture because, although these may have an important influence on fire behaviour, globally comprehensive data sets are not available."

Line 173 “As model benchmarking techniques become more sophisticated it would be beneficial to better evaluate the datasets the models are compared against to ensure the models are being benchmarked appropriately” Please, delete or rephrase (shorten)

We have substantially shortened and modified this paragraph based on your feedback as well as the comments by the first reviewer (see response to reviewer 1 for detailed information regarding the changes made).

Line 175-180. I think you should move formulae of NME from Supplementary back to the main text. You write in Supplementary that you applied NME for areas burnt, but it is clear from Table 3 and S1 that you apply the same metrics for other variables for benchmarking, please, correct.

Agreed. We have put the formula of NME to the main manuscript. Additionally, we have substantially rewritten the section on the benchmarking metrics to make the procedure clearer, including clarifying that NME is used for all the spatial variables.

As well Table 3 is quite difficult to read, why not to use semaphore colors (not so good- red, OK –yellow, good –green) or any other color scheme?

We thank the reviewer for his suggestion. As the objective of this table is to present the numerical scores so that readers will be able to judge differences between the models based on the actual scores, we keep these in the table as well. However, we have now added background colors to identify large-scale differences between the scores. We will modify the caption to this table as follow: “Cell are coloured blue if the benchmarking score is lower than both null models, yellow if lower than 1 null model and red when higher than both null models.”.

Figure 1p. Performance in fire emissions by the models is the worst from all variables. How you can explain it? How reliable is observation data set? Please, make more explanation in Discussion part.

We discuss why the scores for emissions are worse than the scores for other aspects of fire already in the final paragraph of the Discussion. As we state there, errors in the emissions are a product of errors in burnt area, simulated biomass and combustion completeness. It is therefore natural that they are more difficult to predict accurately.

Line 219 CLM (NME: 0.63-0.80) and ORCHIDEE-SPITFIRE (0.70-0.73) are the best performing models. What makes these models to be the best in burnt areas description (for CLM as I understood it is related to cropland fires, what about ORCHIDEE-SPITFIRE ?

We focus on model structure and processes in this manuscript, and only focus on an individual model if it is the only one representing a certain process (e.g. CLM and cropland fires). In this case, both models prescribe vegetation cover, and we have shown that vegetation *is one of the variables influencing performance in simulated burnt area. In fact, both models have the best scores at simulated vegetation carbon.*

Line 326 to 334 ". . . the overall difference between the models (. . .JSBACH, LPJ-GUESS and ORCHIDEE..) reflect feedbacks between the fire and vegetation modules " what are these feedbacks? Where lays difference in their descriptions of these three models (in DGVMs)?

Fires in DGVMs combust biomass & grass and kill trees, hence exerting a strong impact on vegetation dynamics, which themselves drive fire occurrence and characteristics. The three models mentioned are completely different vegetation models. While fire can impact a large range of processes in LPJ-GUESS, ranging from carbon stocks, over distribution and structure. This is less the case for JSBACH and ORCHIDEE as these do not represent vegetation structure and prescribe vegetation distribution. We provide a very detailed description of each fire model in the FireMIP protocol paper (Rabin et al., 2017) and we also provide the key reference providing the description of each vegetation model. We feel that the description of the structure of each vegetation model is outside the scope of this manuscript.

Lines 345-349 "there is a positive relationship between simulated burnt area scores and the seasonal concentration of GPP (R2 = 0.30-0.84) and, to a lesser extent, the seasonal phase of GPP (R2 = 0.09-0.24). This supports the idea that seasonal vegetation production and senescence, which have an important influence on fuel loads, drive the interactions between vegetation and fire within each model" – I doubt this statement. It is more likely that similar dynamics of burnt areas and the seasonal concentration of GPP/ the seasonal phase of GPP are related to dependence of both areas burnt and GPP variables from soil moisture in the fire models and DGVMs. Please, either prove, or delete this paragraph.

We agree that soil moisture, vegetation productivity and fire occurrence are strongly linked to each other and it could be difficult to separate their individual influences. Some recent fire models explicitly represent fuel moisture (e.g. SPITFIRE; Thonicke et al., 2010), with the objective of diagnosing the role of fuel moisture on fire spread. However, soil moisture is only one of the multiple variables which drive fuel moisture, which results in a partial disconnection between soil moisture and fuel moisture in the models. Furthermore, the influence of differences in soil moisture dynamics between models is likely to be small in our experiments because the climate inputs controlling this (precipitation, temperature) were specified to be the same. Thus, we doubt that this feature is solely induced by soil moisture. Furthermore, in addition to this relationship between seasonal vegetation production and burnt area, we provide multiple other indicators that vegetation status impacts the performance of the fire module. This was also a conclusion from a previous study on FireMIP outputs (Forkel et al., 2019). Hence, we believe that there is enough evidence to support our statement that our results stress the importance of the interactions between vegetation and fire within each model.

References

Forkel, M., Andela, N., Harrison, S. P., Lasslop, G., van Marle, M., Chuvieco, E., Dorigo, W., Forrest, M., Hantson, S., Heil, A., Li, F., Melton, J., Sitch, S., Yue, C., and Arneth, A.: Emergent relationships with respect to burned area in global satellite observations and fire-enabled vegetation models, Biogeosciences, 16, 57-76, 10.5194/bg-16-57-2019, 2019.

Hantson, S., Arneth, A., Harrison, S. P., Kelley, D. I., Prentice, I. C., Rabin, S. S., Archibald, S., Mouillot, F., Arnold, S. R., Artaxo, P., Bachelet, D., Ciais, P., Forrest, M., Friedlingstein, P., Hickler, T., Kaplan, J. O., Kloster, S., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Melton, J. R., Meyn, A., Sitch, S., Spessa, A., van der Werf, G. R., Voulgarakis, A., and Yue, C.: The status and challenge of global fire modelling, Biogeosciences, 13, 3359-3375, 10.5194/bg-13-3359-2016, 2016.

Rabin, S. S., Melton, J. R., Lasslop, G., Bachelet, D., Forrest, M., Hantson, S., Kaplan, J. O., Li, F., Mangeon, S., Ward, D. S., Yue, C., Arora, V. K., Hickler, T., Kloster, S., Knorr, W., Nieradzik, L., Spessa, A., Folberth, G. A., Sheehan, T., Voulgarakis, A., Kelley, D. I., Prentice, I. C., Sitch, S., Harrison, S., and Arneth, A.: The Fire Modeling Intercomparison Project (FireMIP), phase 1: experimental and analytical protocols with detailed model descriptions, Geosci. Model Dev., 10, 1175-1197, 10.5194/gmd-10-1175-2017, 2017.

Thonicke, K., Spessa, A., Prentice, I. C., Harrison, S. P., Dong, L., and Carmona-Moreno, C.: The influence of vegetation, fire spread and fire behaviour on biomass burning and trace gas emissions: results from a process-based model, Biogeosciences, 7, 1991-2011, 10.5194/bg-7-1991-2010, 2010.

List of all relevant changes made

1) Updated title in line with reviewer suggestion.
2) Included an extra sentence regarding the performance of the two worst performing models, next to some minor rewriting.
3) 2 minor changes of the introduction in line with reviewer suggestions.
4) Indicated clearer what information can be found in the FireMIP protocol manuscript, based on reviewer questions.
5) Added to the methods a clarification regarding the time period used for analysis, in line with reviewer comments.
6) Section methods section 2.2: clarified what where the criteria for dataset inclusion in the benchmarking exercise, in line with reviewer comments.
7) An overall adaptation of the benchmarking description in the methods to improve clarity, as requested by reviewers.
8) Figure 2: added units in figure caption.
9) Correct two small mistakes in the results section 3.1 which where pointed out by the reviewer.
10) Adapted two sentences to the results section 3.3 regarding the performance of the two worst performing models, as requested by the reviewers. We now also indicate in this section that fires can influence the seasonality in vegetation productivity, as requested by the reviewers.
11) We have added and clarified multiple aspects of the discussion section in line with the multiple suggestions of both reviewers, including a final paragraph regarding the use of benchmarking scores for model ranking.
12) Table 2: Values for JULES-INFERNO where from before correcting a small bug in the code when reading in the JULES-INFERNO landcover file and we have now corrected these.
13) Table 3: we now included background colors to the table to easier differentiate between well and worse performing models. We chose to use blue instead of green to take into account colourblind readers.

Quantitative assessment of fire and vegetation properties in simulations with fire-enabled vegetation models from the Fire Model Intercomparison Project

Deleted: historical

Stijn Hantson[1,2], Douglas I. Kelley[3], Almut Arneth[1], Sandy P. Harrison[4], Sally Archibald[5], Dominique Bachelet[6], Matthew Forrest[7], Thomas Hickler[7,8], Gitta Lasslop[7], Fang Li[9], Stephane Mangeon[10,a], Joe R. Melton[11], Lars Nieradzik[12], Sam S. Rabin[1], I. Colin Prentice[13], Tim Sheehan[6], Stephen Sitch[14], Lina Teckentrup[15,16], Apostolos Voulgarakis[10], Chao Yue[17].

[1] Karlsruhe Institute of Technology, Institute of Meteorology and Climate research, Atmospheric Environmental Research, Garmisch-Partenkirchen, Germany.
[2] Geospatial Data Solutions Center, University of California Irvine, CA 92697 Irvine, USA.
[3] UK Centre for Ecology & Hydrology, Wallingford OX10 8BB, UK.
[4] School of Archaeology, Geography and Environmental Sciences, University of Reading, Reading, UK.
[5] Centre for African Ecology, School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Private Bag X3, WITS, Johannesburg, 2050, South Africa.
[6] Biological and Ecological Engineering, Oregon State University, Corvallis, OR 97331, USA.
[7] Senckenberg Biodiversity and Climate Research Institute (BiK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany.
[8] Institute of Physical Geography, Goethe-University, Altenhöferallee 1, 60438 Frankfurt am Main, Germany.
[9] International Center for Climate and Environmental Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China.
[10] Department of Physics, Imperial College London, London, UK.
[11] Climate Research Division, Environment and Climate Change Canada, Victoria, BC V8W 2Y2, Canada
[12] Institute for Physical Geography and Ecosystem Sciences, Lund University, 22362 Lund, Sweden.
[13] AXA Chair of Biosphere and Climate Impacts, Grand Challenges in Ecosystem and the Environment, Department of Life Sciences and Grantham Institute – Climate Change and the Environment, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, UK.
[14] College of Life and Environmental Sciences, University of Exeter, Exeter EX4 4RJ, UK.
[15] ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, NSW, Australia.
[16] Climate Change Research Center, University of New South Wales, Sydney, NSW 2052, Australia.
[17] Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, 91198 Gif-sur-Yvette, France.
[a] now at: Data 61, CSIRO, Brisbane, Australia

Correspondence to: Stijn Hantson (hantson.stijn@gmail.com)

Deleted: and

Abstract. Global fire-vegetation models are widely used to assess impacts of environmental change on fire regimes and the carbon cycle, and to infer relationships between climate, land use, and fire. However, differences in model structure and parameterizations, in both the vegetation and fire components of these models, could influence overall model performance, and to date there has been limited evaluation of how well different models represent various aspects of fire regimes. The Fire Model Intercomparison Project (FireMIP) is coordinating the evaluation of state-of-the-art global fire models, in order to improve projections of fire characteristics and fire impacts on ecosystems and human societies in the context of global environmental change. Here we perform a systematic evaluation of historical simulations made by nine FireMIP models to quantify their ability to reproduce a range of fire and vegetation benchmarks. The FireMIP models simulate a wide range in global annual total burnt area (39-536 Mha), and global annual fire carbon emission (0.91-4.75 Pg C a^{-1}) for modern conditions (2002-2012), but most of the range in burnt area is within observational uncertainty (345-468 Mha). Benchmarking scores indicate that seven out of nine FireMIP models are able to represent the spatial pattern in burnt area. The models also reproduce the seasonality in burnt area reasonably well but struggle to simulate fire season length and are largely unable to represent inter-annual variations in burnt area. However, models that represent cropland fires see improved simulation of fire seasonality in the northern hemisphere. The three FireMIP models which explicitly simulate individual fires are able to reproduce the spatial pattern in number of fires, but fire sizes are too small in key regions and this results in an underestimation of burnt area. The correct representation of spatial and seasonal patterns in vegetation appears to correlate with a better representation of burnt area. 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.

Deleted: with the aim of improving

Deleted: regime characteristic

Deleted: under

Deleted: in order

Deleted: While some

Deleted: no model

1 Introduction

Fire is a crucial ecological process that affects vegetation structure, biodiversity, and biogeochemical cycles in all vegetated ecosystems (Bond et al., 2005; Bowman et al., 2016) and has serious impacts on air quality, health, and economy (e.g. Bowman et al., 2009; Lelieveld et al., 2015; Archibald et al., 2013). In addition to naturally occurring wildland fires, fire is also used as a tool for pasture management and to remove crop residues. Because fire affects a large range of processes within the Earth system, modules which simulate burnt area and fire emissions are increasingly included in dynamic global vegetation models (DGVMs) and Earth System Models (ESMs) (Hantson et al., 2016; Kloster and Lasslop, 2017; Lasslop et al., 2019). However, the representation of both lightning-ignited fires and anthropogenic fires (including cropland fires) varies greatly in global fire models. This arises due to the lack of a comprehensive understanding of how fire ignitions, spread, and suppression are affected by weather, vegetation, and human activities, as well as the relative scarcity of long-term, spatially resolved data on the drivers of fires and their interactions (Hantson et al., 2016). As a result, model projections of future fire are highly uncertain (Settele et al., 2014; Kloster and Lasslop, 2017). Since vegetation mortality – including fire-related death – is one determinant of carbon residence time in ecosystems (Allen et al., 2015), differences in the representation of fire in DGVMs or ESMs also contributes to the uncertainty in trajectories of future terrestrial carbon uptake (Ahlström et al., 2015; Friend et al., 2014; Arora & Melton, 2018). Improved projections of wildfires and anthropogenic fires, their impact on ecosystem properties, and their socio-economic

Deleted: wildfires

impact will therefore support a wide range of global environmental change assessments, as well as the development of strategies for sustainable management of terrestrial resources.

Although individual fire-enabled DGVMs have been evaluated against observations, comparisons of model performance under modern-day conditions tend to focus on a limited number of fire-related variables or specific regions (e.g. French et al., 2011; Wu et al., 2015; Ward et al., 2016; Kloster and Lasslop, 2017). Such comparisons do not provide a systematic evaluation of whether different parameterizations or levels of model complexity provide a better representation of global fire regimes than others. Likewise, none of the Coupled Model Intercomparison Projects that have been initiated to support the IPCC process (CMIP: Taylor et al., 2012; Eyring et al., 2016) focuses on fire, even though several of the CMIP models simulate fire explicitly. The Fire Model Intercomparison Project (FireMIP) is a collaborative initiative to systematically evaluate state-of-the-art global fire models (Hantson et al., 2016; Rabin et al., 2017).

The FireMIP initiative draws on several different types of simulations, including a baseline historical simulation (1700-2013 CE) and sensitivity experiments to isolate the response of fire regimes to individual drivers, as well as simulations in which fire is deliberately excluded (Rabin et al., 2017). While the sensitivity and exclusion experiments provide valuable insights into model behaviour (Teckentrup et al., 2019; Li et al., 2019), the baseline historical simulation provides an opportunity to assess how well the models simulate modern conditions. Model-model differences could reflect differences in the treatment of fire, of ecosystem processes, or how fire interacts with other aspects of the land surface in an individual model. Evaluation of the baseline simulations needs therefore to include evaluation of ecosystem processes and diagnosis of interactions between simulated vegetation and fire.

Systematic model evaluation can also serve another purpose. The analysis of future climate and climate impacts is often based on results from climate and impact model ensembles (e.g. Kirtman et al., 2013; Collins et al., 2013; Warszawski et al. 2013) and these ensembles are also being used as a basis for impact assessments (e.g. Settele et al., 2014; Hoegh-Guldberg et al., 2019). However, there is increasing dissatisfaction with the idea of using the average behaviour of model ensembles without accounting for the fact that some models are less reliable than others (Giorgi and Mearns 2002; Knutti, 2010; Parker et al., 2013) and many have called for "the end of model democracy" (e.g. Held, 2005; Knutti, 2010). Although there is still considerable discussion about how to constrain models using observations, and then how to combine and possibly weight models depending on their overall performance or performance against a minimum set of specific criteria (e.g. Eyring et al., 2005; Tebaldi et al., 2005; Gleckler et al., 2008; Weigel et al., 2008; Santer et al., 2009; Parker, 2013; Abramowitz et al., 2019), it is clear that results from systematic evaluations are central to this process.

A number of papers have examined specific aspects of the FireMIP baseline simulations. Andela et al. (2017) showed that the FireMIP models do not reproduce the decrease in global burnt area over the past two decades inferred from analysis of version 4s of the Global Fire Emission Database (GFED4s) data product. In fact, four of the models show an increase in burnt area over the period 1997-2014. Although the remaining five models show a decrease, their mean decrease is only about one tenth of the observed rate ($-0.13 \pm 0.56\%$ yr^{-1}, compared to the observed trend of $-1.09 \pm 0.61\%$ yr^{-1}). However, the observed global decline of burnt area derived from satellite data is strongly dominated by African savanna ecosystems, the spatial pattern of trends is very heterogeneous, and the satellite record is still very short, which raises issues about the robustness of these trends (Forkel et al., 2019b). Li et al. (2019) compared modelled and satellite-based fire emissions and concluded that most FireMIP models fall within the current range of observational uncertainty. Forkel et al. (2019a) compared the emergent relationships between burnt area and multiple potential drivers of fire behaviour, including human caused ones, as seen in observations and the FireMIP models. They show that, although all of the models capture the observed emergent relationships with climate variables, there are large differences in their ability to capture vegetation-related relationships. This is underpinned by a regional study using the FireMIP models over China, which showed that there are large differences in simulated vegetation biomass, and hence in fuel loads, between the models (Song et al., 2019). These results make a focus on benchmarking both simulated fire and vegetation particularly pertinent. Forkel et al. (2019a) showed that some of the FireMIP models, specifically those that include a relatively strong fire suppression associated with human activities (Teckentrup et al., 2019), were able to reproduce the emergent relationship with human population density. However, the treatment of the anthropogenic influence on burnt area has been identified as a weakness in the FireMIP models (Andela et al., 2017; Teckentrup et al., 2019; Li et al., 2019; Forkel et al., 2019a), mainly due to a lack of process understanding.

In this paper, we focus on quantitative evaluation of model performance using the baseline historical simulation and a range of vegetation and fire observational datasets. We identify (i) common weaknesses of the current generation of global fire-vegetation models, (ii) factors causing differences between the models, and (iii) discuss the implications for future model development.

Deleted: priorities

2 Methods

2.1 Model Simulations

The baseline FireMIP simulation is a transient experiment starting in 1700 CE and continuing to 2013 (see Rabin et al. (2017) for description of the modelling protocol and the sources of the input data for the experiments). Models were spun up until carbon stocks were in equilibrium for 1700 CE conditions (equilibrium was defined as <1% change over a 50 year time period for the slowest carbon pool in each grid cell) using land use and population density for 1700 CE, CO_2 concentration for 1750 CE, and recycling climate and lightning data from 1901-1920 CE. Although the experiment is fully transient after 1700 CE, annually varying values of all these forcings are not available until after 1900 CE. Climate, land use, population and lightning were regridded to the native grid of each model. Global fire-vegetation models ran with either dynamic or prescribed natural vegetation (Table 1), but all used observed time-evolving cropland and pasture (if simulated) distribution.

Deleted: .,

Deleted:).

Nine coupled fire-vegetation models have performed the FireMIP baseline experiments. The models differ in complexity, representation of human impact and vegetation dynamics, and spatial and temporal resolution (Table 1). A detailed description of each model is given in Rabin et al. (2017). Most of the models ran simulations for the full period 1700-2013, but CLASS-CTEM, JULES-INFERNO, MC2 and CLM simulated 1861-2013, 1700-2012, 1902-2009 and 1850-2013 respectively. This slight deviation from the protocol does not affect the results all but 1 model presented here as we only analyse data for present-day period (2002-2012). For MC2 the 2002-2009 time period was used for analysis which might influence the results for this model.

Deleted: .

2.2) Model Evaluation and Benchmarking

Model performance was evaluated using site-based and remotely sensed global data sets of fire occurrence, fire-related emissions and vegetation properties (Figure 1; Figure S1). We include vegetation variables (e.g. GPP, NPP, biomass, LAI) because previous analyses have indicated that they are critical for simulating fire occurrence and behaviour (Forkel et al., 2019a; Teckentrup et al., 2019) and there are global data sets available. We did not consider parameters such as soil or litter moisture because, although these may have an important influence on fire

Deleted: vegetation properties,

Deleted: and

behaviour, globally comprehensive data sets are not available. We used multiple datasets as targets for variables where they were available in order to take into account observational uncertainty. We used the GFED4 (Giglio et al., 2013), GFED4s (Randerson et al., 2012), MCD45 (Roy et al., 2008), FireCC4.0 (Alonso-Canas and Chuvieco, 2015) and FireCCI5.1 (Chuvieco et al., 2018) burnt area datasets; the Global Fire Assimilation System (GFAS) fire emissions (Kaiser et al., 2012); fire size and number from Hantson et al. (2015); site-level net primary productivity (NPP) and gross primary productivity (GPP) from Luyssaert et al. (2007), as well as NPP from the Ecosystem Model/Data Intercomparison (EMDI; Olson et al., 2001), from Michaletz et al. (2014) and upscaled fluxnet GPP data (Jung et al., 2017; Tramontana et al., 2017); aboveground vegetation biomass from Carvalais et al. (2014) and Avitabile et al. (2016); and Leaf Area Index (LAI) from MODIS MCD15 (Myneni et al., 2002) and AVHRR (Claverie et al., 2016). A complete description of the reference data sets used for benchmarking is given in the Supplementary Information S1 and all datasets are plotted in Figure S1. For comparison and application of the benchmark metrics, all the target datasets and model outputs were resampled to a 0.5° grid. Although some models were run at a coarser resolution, the spatial resolution at which the benchmarking was performed had only a limited impact on the scores (Figure S2), which does not affect conclusions drawn here. Each model was compared to each reference dataset except in the few cases where the appropriate model output was not provided (e.g. LAI in ORCHIDEE, GPP in MC2). Only the models which incorporate the SPITFIRE fire module provided fire size and number results.

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 assessment of the impact 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 in calculating the benchmarking scores.

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

Deleted: (Figure 1; Figure S1).

Deleted: While we use multiple datasets to attempt to quantify observational uncertainty, this only addresses one area of observational uncertainty. Most observation-based datasets don't provide estimates of measurement error and uncertainty. As well, most datasets haven't been independently validated and could suffer from methodological issues affecting quality. Therefore, large uncertainties still remain for most variables, even well-established ones such as burnt area (e.g. Roteta et al., 2019). When uncertainties of datasets are known one can take this into account by weight the model scores less for the less trustworthy datasets (e.g. Collier et al. 2018). However, as the uncertainty of most reference datasets are currently unknown, we haven't attempted that here and instead treat each dataset equally. As model benchmarking techniques become more sophisticated it would be beneficial to better evaluate the datasets the models are compared against to ensure the models are being benchmarked appropriately.
We use the normalised mean error (NME), which is the mean absolute difference between a simulated and observed variable normalised by the observational variance, to

Deleted: each model's

Deleted: annual

Deleted: burnt area and the other variables that influence burnt area. We therefore

Moved (insertion) [1]

Deleted: reference datasets (See Supplementary Information S2) because NME can be sensitive to the simulated magnitude and this furthermore limits

Deleted: Seasonality comparisons were conducted in two parts:

Deleted: ,

Deleted: , of the season.

Deleted: in a given

Deleted: or simulated cell

Deleted: burnt area throughout

Deleted: burnt area

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.

Deleted: fire

Deleted: equally spread

Deleted: 1 when all burning occurs in the same

Deleted: , and phases

Deleted: Three models simulate or provided only annual burnt area …

Deleted:) and thus the seasonality benchmarks could not be calculated for …

Deleted: models

Deleted:

Deleted: NME and MPD are both proportional to mean absolute errors, so the smaller the value the better the model performance.

Moved up [1]: A score of 0 represents a perfect match to observations.

Deleted: NME has no upper bound, whereas MPD has a maximum value of 1 when all cells have a maximum phase mismatch of 6 months. …

Deleted: further

Deleted:): the

Deleted: the mean value of the observations to the observations; and the…

Deleted: to the observations

Deleted: A detailed description of the benchmarking metrics is given in the Supplementary Information S2.

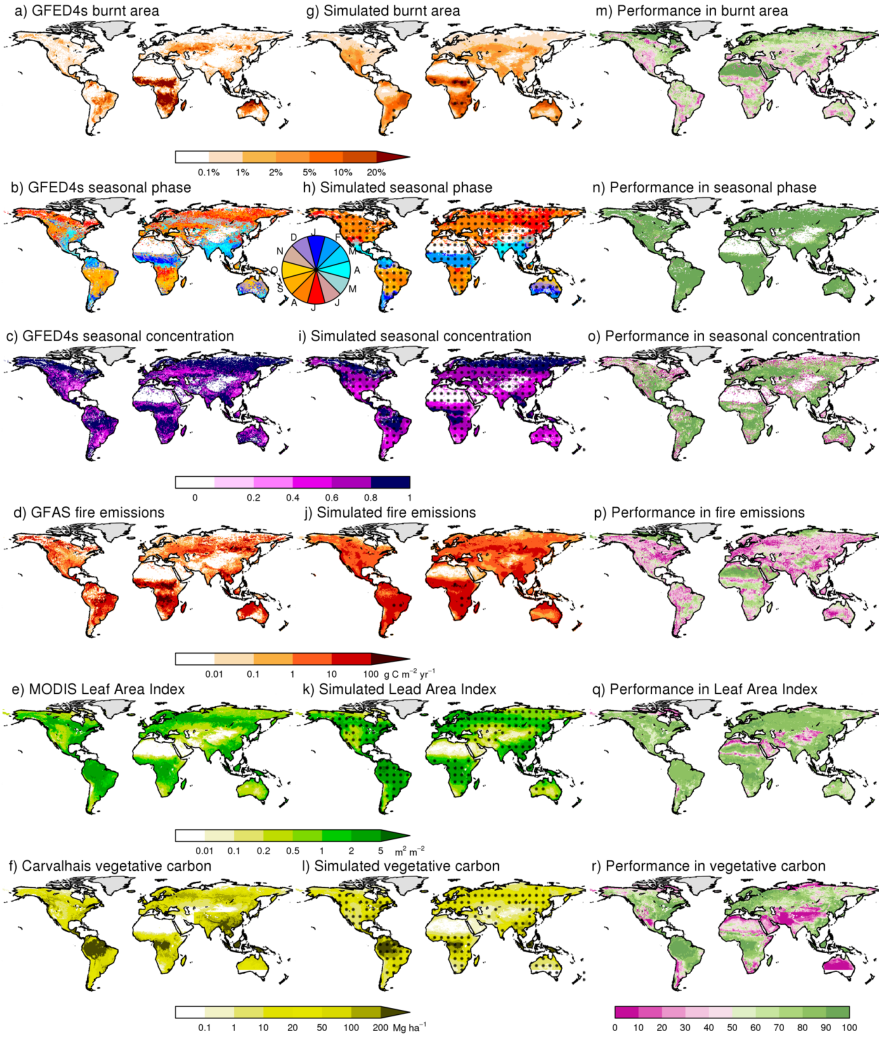
a) GFED4s burnt area
g) Simulated burnt area
m) Performance in burnt area
0.1% 1% 2% 5% 10% 20%
b) GFED4s seasonal phase
h) Simulated seasonal phase
n) Performance in seasonal phase
J F M A M J J A S O N D
c) GFED4s seasonal concentration
i) Simulated seasonal concentration
o) Performance in seasonal concentration
0 0.2 0.4 0.6 0.8 1
d) GFAS fire emissions
j) Simulated fire emissions
p) Performance in fire emissions
0.01 0.1 1 10 100 g C m^{-2} yr^{-1}
e) MODIS Leaf Area Index
k) Simulated Lead Area Index
q) Performance in Leaf Area Index
0.01 0.1 0.2 0.5 1 2 5 m^2 m^{-2}
f) Carvalhais vegetative carbon
l) Simulated vegetative carbon
r) Performance in vegetative carbon
0.1 1 10 20 50 100 200 Mg ha^{-1}
0 10 20 30 40 50 60 70 80 90 100

Figure 1: Reference datasets, the mean of all models, and the % of models for which the estimate falls within 50-200% of the (mean) reference data are presented for a set of fire relevant variables. Results for the following variables are given: a) fraction burnt area; b) seasonal timing of burnt area (as measured by mean phase); c) burnt area season length (as measured by seasonal concentration); d) fire C emissions (g C m^{-2} yr^{-1}); e) vegetation carbon (Mg/ha); and f) Leaf Area Index (LAI) (m^2/m^2). Stippling in the 2nd column indicates where variance between models is less than the FireMIP model ensemble mean. Purple in the 3rd column indicates cell where the majority of the FireMIP models produce poor simulations of the variable, while green areas indicate that the majority of the FireMIP models perform well for that aspect of the fire regime.

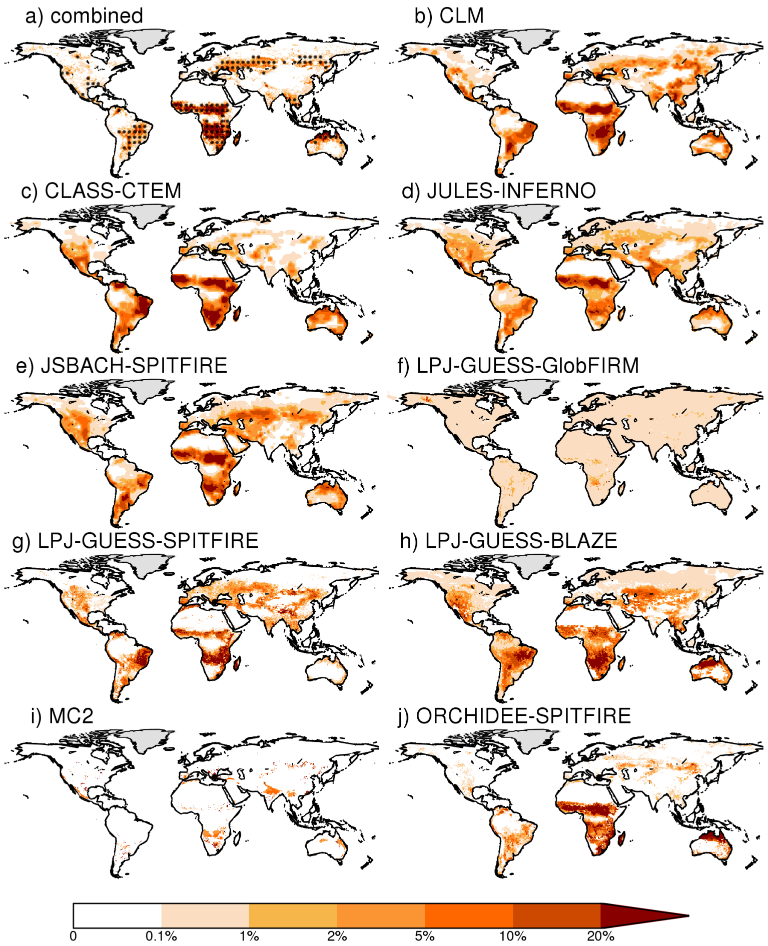

Figure 2: Simulated versus observed burnt fraction (% yr^{-1}) for the present day (2002-2012), where "combined" indicates the mean of the different burnt area datasets considered. Stippling indicates where variance between burnt area datasets is less than the observed ensemble mean.

Deleted: in

3 Results

3.1 Modern day model performance: burnt area and fire emissions

The simulated modern (2002-2012) total global annual burnt area is between 39 and 536 Mha (Table 2). Most of the FireMIP models are within the range of the remotely sensed observed burnt area (354 to 468 Mha a^{-1}). With the exception of MC2 and LPJ-GUESS-GlobFIRM, the models realistically capture the spatial patterns in burnt area (Figures 1 & 2) and perform better than either of the null models irrespective of the reference burnt area dataset (Table 3). CLM (NME: 0.63-0.80) and ORCHIDEE-SPITFIRE (0.70-0.73) are the best performing models. All the FireMIP models correctly simulate most burnt area in the tropics (24-466 Mha a^{-1}) compared to observed values in the range 312-426 Mha a^{-1} (Table 2). The simulated contribution of tropical fires to global burnt area is in the range of 56% to 92%, with all models except ORCHIDEE-SPITFIRE simulating a lower fraction than observed (89-93%). This follows from FireMIP models tending to underestimate burnt area in Africa and Australia, although burnt area in South American savannas is usually overestimated (Table 2). All of the FireMIP models, except LPJ-GUESS-GlobFIRM, capture a belt of high burnt area in central Eurasia. However, the models overestimate burnt area across the extratropics on average by 180% to 304%, depending on the reference burnt area dataset. This overestimation largely reflects the fact that the simulated burnt area over the Mediterranean basin and western USA is too large (Table 2, Figure 2).

The FireMIP models that include a sub-annual time-step for fire calculations (CLM, CLASS-CTEM, JULES-INFERNO, JSBACH-SPITFIRE, LPJ-GUESS-SPITFIRE, ORCHIDEE-SPITFIRE) generally reproduce the seasonality of burnt area (Figure 3), particularly in the tropics. The models capture the timing of the peak fire season reasonably well, with all of the models performing better than mean null model for seasonal phase in burnt area (Table 3). The models also frequently perform better than the random null model, with all models performing better against GFED4. However, all of the FireMIP models perform worse than mean null model for seasonal concentration of burnt area, independent of the reference burnt area dataset. The observations show a unimodal pattern in burnt area in the tropics, peaking between November through February in the northern tropics and between June through October in the southern tropics (Figure 3). The models also show a unimodal pattern in both regions. However, all the FireMIP models except ORCHIDEE-SPITFIRE show a ~2-month delay in peak burnt area in the northern tropics, and the period with high burnt area is also less concentrated than observed. Some models (ORCHIDEE-SPITFIRE, LPJ-GUESS-SPITFIRE) estimate peak burnt area ~1-2 months too early in the southern tropics, while others simulate a peak ~1 month too late (JULES-INFERNO, CLM, CLASS-CTEM) or have a less concentrated peak (JSBACH-SPITFIRE, JULES-INFERNO) than observed. The seasonality of burnt area in the northern extratropics shows a peak in spring and a second peak in summer. Only CLM reproduces this double peak, while all of the other FireMIP models show a single summer peak. Most of the models simulate the timing of the summer peak well. The only exception is LPJ-GUESS-SPITFIRE, which simulates the peak ~2-3 months too late. The observations show no clear seasonal pattern in burnt area over the southern extratropics, although the most prominent peak occurs in December and January. All the FireMIP models, except LPJ-GUESS-SPITFIRE, reproduce this mid-summer peak. LPJ-GUESS-SPITFIRE shows little seasonality in burnt area in this region.

Deleted: both

Deleted: models

Deleted: both

Deleted: models

The FireMIP models have problems representing IAV in global burnt area, with some models (CLASS-CTEM, MC2) worse than the random model and most models performing worse than the mean for most of the target data sets (Table 3). However, there is considerable uncertainty in the observed IAV in burnt area (Figure 4), and the scores are therefore dependent on the reference dataset considered, with generally worse scores for FireCCI5.1 and GFED4s compared to the other datasets. Observational uncertainty is most probably underestimated as the burnt area products are not independent, since they all rely on MODIS satellite imagery. Despite the failure to reproduce IAV in general, most of the models show higher burnt area in the early 2000s and a low in 2009-2010 after which burnt area increased again (Figure 4).

The spatial patterns in simulated fire-related carbon emissions are in line with the reference data, with most FireMIP models except LPJ-GUESS-GlobFIRM, MC2 and LPJ-GUESS-SPITFIRE performing better than the mean null model. CLM, JULES-INFERNO and JSBACH-SPITFIRE are the best performing models with NME scores < 0.8. Seasonality in fire emissions mimics the results for burnt area with good scores for seasonal phase, but all models perform worse than the mean null model for seasonal concentration. CLM is the only FireMIP model to explicitly include peatland, cropland and deforestation fires, which contribute 3%, 3% and 20% respectively of the global total emissions annually (van der Werf et al., 2010), but it nevertheless does not perform better than JULES-INFERNO and JSBACH-SPITFIRE in representing the spatial pattern of fire carbon emissions.

Only three models (JSBACH-SPITFIRE, LPJ-GUESS-SPITFIRE, ORCHIDEE-SPITFIRE) provided information about simulated numbers and size of individual fires. All three models performed better than the mean null model in representing the spatial pattern in number of fires but worse than the mean model for fire size (Table 3). While the spatial pattern in simulated fire number is in agreement with observations over large parts of the globe, models tend to overestimate fire numbers in dryland areas such as Mexico and the Mediterranean basin (Figure 5). None of the three models simulate cropland fires and so they do not capture the high number of cropland fires (Hall et al., 2016) in central Eurasia (Table 2). Models simulate smaller fires than observed in areas where burnt area is large and where models tend to underestimate burnt area, especially in the African savanna regions (Figure 5).

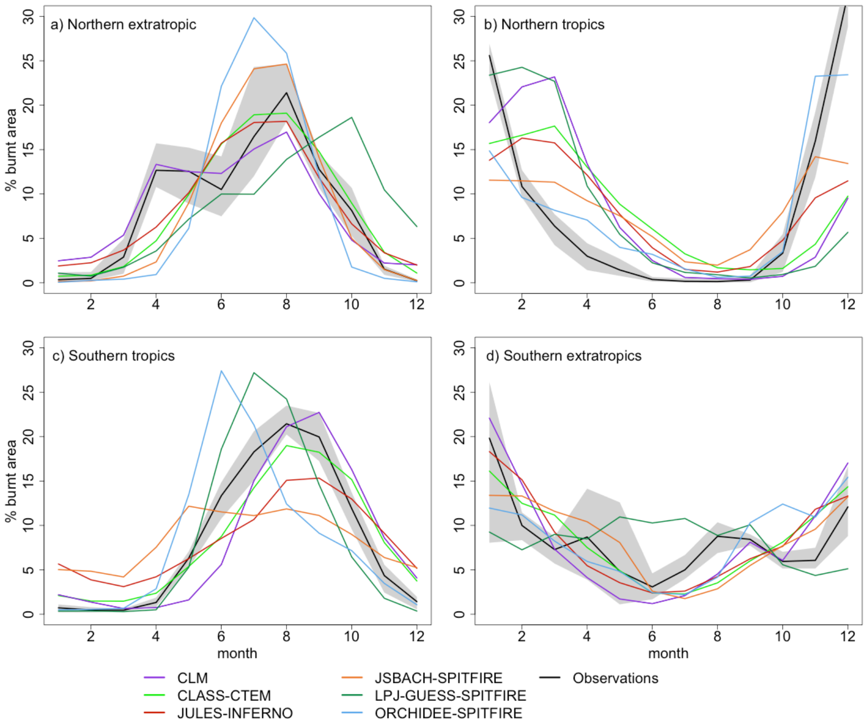

Figure 3: Simulated and observed seasonality (2002-2012) of burnt area (% of annual burnt area per month) for a) northern extratropics (> 30°N), b) northern tropics (0-30°N), c) southern tropics (0-30°S) and d) southern extratropics (> 30°S). The mean of all the remotely sensed burnt area datasets is shown as a black line, with the minimum and maximum range shown in light grey.

3.2. Present day model performance: Vegetation properties

Fire spread and hence burnt area is strongly influenced by fuel availability, which in turn is affected by vegetation primary production and biomass. Simulated spatial patterns of GPP compare well with estimates of GPP upscaled from Fluxnet data (Jung et al., 2017), with scores (0.39-0.67) considerably better than both null models. However, performance against site-based estimates of GPP (Luyssaert et al., 2007) are considerably poorer (1.09-1.49) and worse than the mean null model. Only LPJ-GUESS-SPITFIRE, LPJ-GUESS-SIMFIRE-BLAZE and ORCHIDEE-SPITFIRE perform better than the random null model. There is no clear relationship between model scores for the two datasets: models performing better when compared to the Jung dataset do not necessarily show a higher score when compared to the Luyssaert GPP dataset. The two GPP datasets are very different: The upscaled FLUXNET dataset is a modelled product but has global coverage (see Supplementary Information S1) while the Luyssaert dataset has local measurements but only at a limited number of sites, largely concentrated across the northern extratropics. Thus, the better match between the FireMIP models and the upscaled FLUXNET dataset may reflect the broader spatial coverage or the fact that climate and landcover data are used for upscaling.

Only the upscaled Fluxnet data provides monthly data and can thus be used to asses GPP seasonality. The FireMIP models are able to represent the seasonal peak timing in GPP, with all models performing better than the mean and random null models. However, models have difficulty in representing the length of the growing season, with the scores for seasonal concentration in GPP (1.08-1.23) above the mean null model but below the random null model for all FireMIP models.

Model performance is better for site-level NPP than site-level GPP. All of the FireMIP models perform better than the mean null model, independent of the choice of reference data set (Table 3), except for CLASS-CTEM against the Luyssaert data set. JULES-INFERNO, JSBACH-SPITFIRE and MC2 are the best-performing models.

The FireMIP models generally capture the spatial pattern in LAI, with all models performing better than the mean null model (0.44-0.81), independent of the reference dataset considered. JULES-INFERNO has the best score for both reference datasets. Although the overall global pattern in LAI is well represented in all the FireMIP models, they have more trouble representing LAI in agricultural areas such as central USA or areas with low LAI such as drylands and mountain areas (Figure 1).

The FireMIP models perform well in representing the spatial pattern carbon in vegetation (Table 3). All nine models perform better than the mean null model, independent of reference dataset, with ORCHIDEE-SPITFIRE having the best scores. Generally, the models are able to simulate carbon in tropical vegetation and the forested regions in the temperate and boreal region reasonably well, but struggle across most dryland systems (Figure 1).

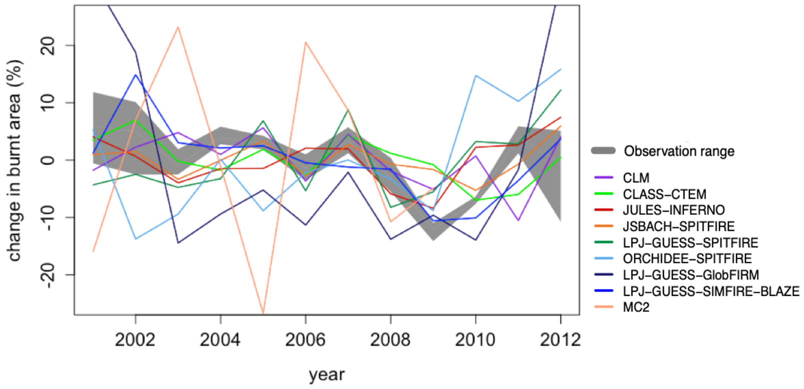

Figure 4: The range in inter-annual variability in burnt area for the years 2001-2012 for all models and burnt area datasets which span the entire time period (GFED4, GFED4s, MCD45, FireCCI51). Results from the individual FireMIP models, as well as the observational minimum-maximum values, are plotted.

3.3 Overall assessment

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. Process-based fire models (see table 1) appear to be slightly better able to represent the spatial pattern in burnt area than empirical models (mean score 0.87 and 0.94 respectively), but this difference is largely the result of including GlobFIRM in the empirical model ensemble; removing this model results in a mean score of 0.87 for these models. The inter-model spread in scores within each group is much larger than the difference between the two types of model. Only one empirical model simulates fire seasonality, but this model performs worse than each of the process-based models, independent of reference dataset considered. There is no difference in the performance of process-based and empirical models with respect to IAV in burnt area, seasonal phase in burnt area or fire emissions.

Deleted: There is no

Deleted: large

The FireMIP simulations include three models in which versions of the same process-based fire module (SPITFIRE) are coupled to different vegetation models. These three models produce a wide range of benchmarking scores for burnt area, with mean benchmarking scores of 0.79, 0.85 and 0.72 for JSBACH, LPJ-GUESS and ORCHIDEE respectively. There are also large differences between these models in terms of other aspects of the fire regime (Table 3). As there are only moderate differences between the different SPITFIRE implementations (Rabin et al., 2017), this suggests that the overall difference between the models reflect interactions between the fire and vegetation modules.

Deleted: feedbacks

Models using prescribed vegetation biogeography (CLM, CLASS-CTEM, JSBACH-SPITFIRE, ORCHIDEE-SPITFIRE) represent the spatial pattern of burnt area better than models with dynamic vegetation (JULES-INFERNO, LPJ-GUESS-SPITFIRE, LPJ-GUESS-GlobFIRM, LPJ-GUESS-SIMFIRE-BLAZE, MC2), with respective mean benchmarking scores across all burnt area data sets of 0.79 and 0.97. This difference is still present even when LPJ-GUESS-GlobFIRM and MC2 are not included (0.90). It seems likely that models using prescribed vegetation biogeography have a better representation of fuel loads and flammability. This can also partially be seen in the positive relationship between the benchmarking scores of vegetation carbon and burnt area spatial patterns for at least the GFED4, FireCCI4.0 and FireCCI5.1 burnt area reference datasets (mean $R^2 = 0.31$, range 0.19-0.38). Areas where the FireMIP models represent vegetation carbon poorly coincide with some of the regions where models have trouble representing the spatial pattern of burnt area such as dryland regions (Figure 1). Although there is no relationship between GPP/NPP and burnt area benchmarking scores, there is a positive relationship between simulated burnt area scores and the seasonal concentration of GPP ($R^2 = 0.30$-0.84) and, to a lesser extent, the seasonal phase of GPP ($R^2 = 0.09$-0.24). This supports the idea that seasonal vegetation production and senescence, which have an important influence on fuel loads, drive the interactions between vegetation and fire within each model. However, since fires combust some of the vegetation, fire occurrence also influences the seasonality in vegetation productivity.

Fire carbon emission benchmarking scores are strongly related to the burnt area performance ($R^2 > 0.85$ for GFED4s and MCD45 and >0.45 for FireCCI4.0 and GFED4). This indicates that simulated burnt area is the main driver of fire emissions, overriding spatial patterns in fuel availability and consumption. However, the benchmarking scores for the spatial pattern in burnt area are better overall than those for fire carbon emissions.

Models that explicitly simulate the impact of human suppression on fire growth or burnt area (CLM, CLASS-CTEM, JSBACH-SPITFIRE, LPJ-GUESS-SIMFIRE-BLAZE) are better at representing the spatial pattern in burnt area compared to models which do not include this effect (0.85 and 0.93 respectively). In the case of the three process-based models (CLM, CLASS-CTEM, JSBACH-SPITFIRE) this is most probably because the spatial pattern in fire size is better represented (Table 3).

CLM is the only model that incorporates cropland fires (Table 1) and it is also the only model which captures the spring peak in burnt area in the northern extratropics associated with crop fires (e.g. Le Page et al., 2010; Magi et al., 2012, Hall et al., 2016). This might also contribute to the good overall score of CLM for spatial pattern of burnt area.

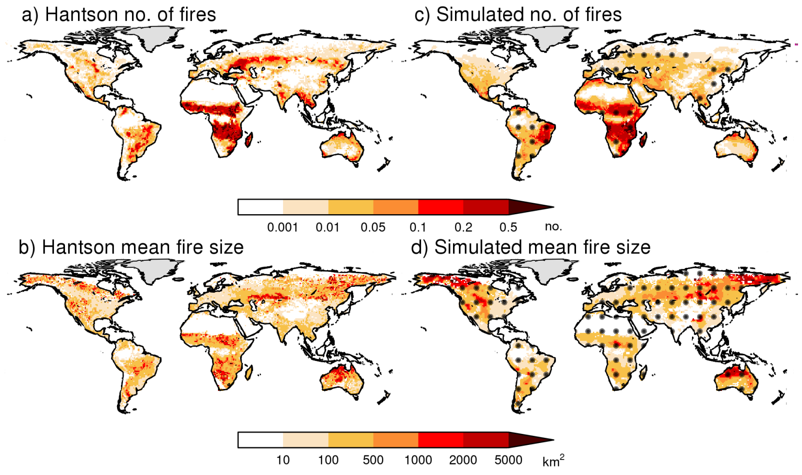
a) Hantson no. of fires
c) Simulated no. of fires
0.001
0.01
0.05
0.1
0.2
0.5
no.
b) Hantson mean fire size
d) Simulated mean fire size
10
100
500
1000
2000
5000
km^2

Figure 5: Reference datasets and mean of three models for number of fires and mean fire size. Model output is adapted so that mean and variance coincide with observations as the total values are not directly comparable (See Supplementary Information S1). Stippling indicates where variance between models is less than the model ensemble mean.

4 Discussion

Deleted: and future model development

There are large differences in the total burnt area between the FireMIP models, with two models (LPJ-GUESS-GlobFIRM and MC2) falling well outside the observed range in burnt area for the recent period. In the case of LPJ-GUESS-GlobFIRM, this is because GlobFIRM was developed before global burnt area products were available, resulting in a general poor performance (Kloster and Lasslop, 2017), in combination with the fact that structural changes were made to the vegetation model without a commensurate development of the fire module. In the case of MC2, this probably reflects the fact that MC2 was developed for regional applications but was applied globally here without any refinement of the fire model. The other FireMIP models used the burned area datasets to develop and tune their models. They therefore capture the global spatial patterns of burnt area reasonably well, although no model simulates the very high burnt area in Africa and Australia causing a general underestimation of burnt area in tropical regions and overestimation in extratropical regions. The analysis of a limited number of models suggests that process-based fire models do not simulate the spatial patterns in fire size well (Table 3). In particular they fail to represent fire size in tropical savannas (Figure 5), most probably because they assume a fixed maximum fire duration of less than one day (Hantson et al., 2016) while savanna fires are often very long-lived (e.g. Andela et al., 2019). Our results suggest that process-based fire models could be improved by a better representation of fire duration. Although none of the FireMIP models simulate multi-day fires, there are fire models that do (e.g. Pfeifer et al., 2013; Le Page et al., 2015) and which could therefore provide a template for future model development. New parameterizations would need to incorporate aspects of natural and anthropogenic landscape fragmentation which limit fire growth (e.g. Pfeifer et al., 2013; Le Page et al., 2015; Kelley et al. 2019). Indeed, our results show that models that include a human limitation on fire growth represent the global spatial pattern in burnt area and fire size better. The recently generated Global Fire Atlas (Andela et al., 2019) includes

Deleted: However,

Deleted: These results indicate that process-based fire models can be improved by a better representation of fire duration.

aspects of the fire behaviour (e.g., fire spread rate and duration), which offer new opportunities to examine and parameterize fire.

Vegetation type and stocks are input variables for the fire models, influencing fire ignition and spread in the process-based models and determining simulated burnt area in the empirical models. The occurrence of fire can, in turn, affect the vegetation type, simulated vegetation productivity (i.e. GPP, NPP) and hence the amount and seasonality of fuel build up. Our results indicate that inter-model differences in burnt area are related to differences in simulated vegetation productivity and carbon stocks. Seasonal fuel build-up and senescence is an important driver of global burnt area. Furthermore, we find that models which are better at representing the seasonality of vegetation production are also better at representing the spatial pattern in burnt area. These results are consistent with the analysis of emergent relationships in FireMIP models, which shows the need to improve processes related to plant production and biomass allocation to improve model performance in simulating burnt area (Forkel et al., 2019a). While there are spatially explicit global estimates regarding carbon stocks in live vegetation, there is limited information about carbon stocks of different fuel types and how these change between seasons and over time (van Leeuwen et al., 2014; Pettinari & Chuvieco, 2016). Furthermore, fuel availability could be substantially affected by livestock density and pasture management (Andela et al., 2017). While improved representation of land management practices could improve the representation of fire, the lack of high-quality fuel availability data currently limits our ability to constrain simulated fuel loads.

The FireMIP models generally do not simulate the timing of peak fire occurrence accurately and tend to simulate a fire season longer than observed. This might be related to the representation of seasonality in vegetation production and fuel build up. However, human activities can also change the timing of fire occurrence (e.g. Le Page et al., 2010; Rabin et al., 2015), and so an improved representation of the human influence on fire occurrence and timing could also help to improve the simulated fire seasonality. The importance of the anthropogenic impact on fire seasonality is especially clear in the northern extratropics (e.g. Archibald et al., 2009; Le Page et al., 2010; Magi et al., 2012), where the only model that explicitly includes crop fires (CLM) is also the only model that shows the bimodal seasonality. Thus, the inclusion of anthropogenic fires could help to improve model simulations. However, this requires a better understanding of how fire is used for land management under different socio-economic and cultural conditions (Pfeiffer et al., 2013; Li et al., 2013).

Global inter-annual variability in burnt area is largely driven by drought episodes in high biomass regions and fuel buildup after periods of increased rainfall in dryland areas (e.g. Chen et al., 2017). Previous analysis has shown that the FireMIP models are relatively good at representing emergent climate-fire relationships (Forkel et al., 2019a); hence it seems plausible that fuel build up and its effect on subsequent burnt area is not well represented in the models and that this is the reason for the poor simulation of IAV in burnt area. This is in line with our findings and the findings of Forkel et al. (2019a) that fire models are not sensitive enough to previous previous-season vegetation productivity.

The spread in simulated global total fire emissions is even larger than for burnt area, but fire emissions largely follow the same spatial and temporal patterns as burnt area (Figure 1, table 3). However, the benchmark scores for emissions are worse than those for burnt area. This reflects the fact that emissions are the product of both errors in simulated vegetation and burnt area. Furthermore, spatial and temporal uncertainties in the completeness of biomass combustion will affect the emissions. While improvements to vegetation and fuel loads are likely to produce more reliable estimates of emissions, an improved representation of the drivers of combustion completeness in models will also be required for more accurate fire emission estimates. Only one of the FireMIP

Deleted: burning processes simulated by these models

Deleted: Inter

Deleted: production

Deleted: This indicates that seasonal fuel build-up and senescence is an important driver of global burnt area.

Deleted: The

Deleted:

Deleted: buildup

models (CLM) includes cropland, peatland, and deforestation fire explicitly, albeit in a rather simple way. Our analyses suggest that this does not produce an improvement in the simulation of either the spatial pattern or timing of carbon emissions. However, given that together these fires represent a substantial proportion of annual carbon emissions, a focus on developing and testing robust parameterisations for these largely anthropogenic fires could also help to provide more accurate fire emission estimates.

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.

Deleted: ¶
¶
¶
¶

Competing interests. The authors declare no competing interest

Author contribution. SH and DK analysed the model results; SH, DB, MF, GL, FL, SM, JRM provided global fire-vegetation model outputs; SH, DK, SPH and AA wrote the manuscript with input from all authors.

Code and data availability. The benchmarking code is archived https://zenodo.org/record/3525235#.XdcBzC2ZOcZ (https://doi.org/10.5281/zenodo.3525235). The FireMIP model output is archived at https://zenodo.org/record/3555562#.Xell3C2ZOcY (DOI:10.5281/zenodo.3551041). Data availability for each reference dataset is provided in the table below:

Variable	Dataset	data access
Burnt area	GFED4	https://www.globalfiredata.org/data.html
	GFED4S	https://www.globalfiredata.org/data.html
	MCD45	Upon request to the author
	FireCCI40	https://geogra.uah.es/fire_cci/
	FireCCI51	https://geogra.uah.es/fire_cci/
Fire emissions	GFAS	https://apps.ecmwf.int/datasets/data/cams-gfas/
Fire size & number	Hantson	https://zenodo.org/record/3564818#.Xelidi2ZOcY
GPP	Luyssaert	Upon request to the author
	Jung	https://www.bgc-jena.mpg.de/geodb/projects/Home.php
NPP	Michaletz	https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fgeb.12685&file=geb12685-sup-0002-suppinfo2.xlsx
	Luyssaert	Upon request to the author
	EMDI	http://gaim.unh.edu/Structure/Intercomparison/EMDI/validationdata2/
LAI	MCD15	http://doi.org/10.5067/MODIS/MCD15A2H.006
	AVHRR	https://www.ncei.noaa.gov/data/avhrr-land-leaf-area-index-and-fapar/access/
Carbon in vegetation	Avitabile	http://lucid.wur.nl/datasets/high-carbon-ecosystems
	Carvalhais	Upon request to the author

Acknowledgments. S. Hantson, S. Rabin, and A. Arneth acknowledge support by the EU FP7 projects BACCHUS (grant agreement no. 603445) and LUC4C (grant ag. No. 603542). This work was supported, in part, by the German Federal Ministry of Education and Research (BMBF), through the Helmholtz Association and its research programme ATMO, and the HGF Impulse and Networking fund. S. P. Harrison acknowledges support under the ERC-funded project GC2.0 (Global Change 2.0: Unlocking the past for a clearer future, grant number 694481). D. Kelley was supported by the UK Natural Environment Research Council through The UK Earth System Modelling Project (UKESM, Grant No. NE/N017951/1). GL was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – 338130981. F. Li is supported by National Natural Science Foundation of China (41475099 and 41875137).

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91-105, 10.5194/esd-10-91-2019, 2019.

Ahlström, A., Xia, J., Arneth, A., Luo, Y., and Smith, B.: Importance of vegetation dynamics for future terrestrial carbon cycling, Environmental Research Letters, 10, 054019, 2015.

Allen, C. D., Breshears, D. D., and McDowell, N. G.: On underestimation of global vulnerability to tree mortality and forest die-off from hotter drought in the Anthropocene, Ecosphere, 6, 1-55, 2015.

Alonso-Canas, I., and Chuvieco, E.: Global burned area mapping from ENVISAT-MERIS and MODIS active fire data, Remote sensing of environment, 163, 140-152, http://dx.doi.org/10.1016/j.rse.2015.03.011, 2015.

Andela, N., Morton, D. C., Giglio, L., Chen, Y., van der Werf, G. R., Kasibhatla, P. S., DeFries, R. S., Collatz, G. J., Hantson, S., Kloster, S., Bachelet, D., Forrest, M., Lasslop, G., Li, F., Mangeon, S., Melton, J. R., Yue, C., and Randerson, J. T.: A human-driven decline in global burned area, Science, 356, 1356-1362, 10.1126/science.aal4108, 2017.

Andela, N., Morton, D. C., Giglio, L., Paugam, R., Chen, Y., Hantson, S., van der Werf, G. R., and Randerson, J. T.: The Global Fire Atlas of individual fire size, duration, speed and direction, Earth Syst. Sci. Data, 11, 529-552, 10.5194/essd-11-529-2019, 2019.

Archibald, S., Roy, D. P., Van Wilgen, B. W., and Scholes, R. J.: What limits fire? An examination of drivers of burnt area in Southern Africa, Global Change Biology, 15, 613-630, 2009.

Archibald, S., Lehmann, C. E. R., Gómez-Dans, J. L., and Bradstock, R. A.: Defining pyromes and global syndromes of fire regimes, Proceedings of the National Academy of Sciences, 110, 6442-6447, 10.1073/pnas.1211466110, 2013.

Arora, V. K., and Melton, J. R.: Reduction in global area burned and wildfire emissions since 1930s enhances carbon uptake by land, Nature Communications, 9, 1326, 10.1038/s41467-018-03838-0, 2018.

Avitabile, V., Herold, M., Heuvelink, G. B. M., Lewis, S. L., Phillips, O. L., Asner, G. P., Armston, J., Ashton, P. S., Banin, L., Bayol, N., Berry, N. J., Boeckx, P., de Jong, B. H. J., DeVries, B., Girardin, C. A. J., Kearsley, E., Lindsell, J. A., Lopez-Gonzalez, G., Lucas, R., Malhi, Y., Morel, A., Mitchard, E. T. A., Nagy, L., Qie, L., Quinones, M. J., Ryan, C. M., Ferry, S. J. W., Sunderland, T., Laurin, G. V., Gatti, R. C., Valentini, R., Verbeeck, H., Wijaya, A., and Willcock, S.: An integrated pan-tropical biomass map using multiple reference datasets, Global Change Biology, 22, 1406-1420, 10.1111/gcb.13139, 2016.

Baccini, A., Goetz, S. J., Walker, W. S., Laporte, N. T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P. S. A., Dubayah, R., Friedl, M. A., Samanta, S., and Houghton, R. A.: Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps, Nature Clim. Change, 2, 182-185, 2012.

Bachelet, D., Ferschweiler, K., Sheehan, T. J., Sleeter, B. M., and Zhu, Z.: Projected carbon stocks in the conterminous USA with land use and variable fire regimes, Global Change Biology, 21, 4548-4560, 2015.

Bond, W. J., Woodward, F. I., and Midgley, G. F.: The global distribution of ecosystems in a world without fire, New Phytologist, 165, 525-538, 2005.

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., Swetnam, T. W., van der Werf, G. R., and Pyne, S. J.: Fire in the Earth System, Science, 324, 481-484, 10.1126/science.1163886, 2009.

Bowman, D. M. J. S., Perry, G. L. W., Higgins, S. I., Johnson, C. N., Fuhlendorf, S. D., and Murphy, B. P.: Pyrodiversity is the coupling of biodiversity and fire regimes in food webs, Philosophical Transactions of the Royal Society of London B: Biological Sciences, 371, 10.1098/rstb.2015.0169, 2016.

Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., u, M., Saatchi, S., Santoro, M., Thurner, M., Weber, U., Ahrens, B., Beer, C., Cescatti, A., Randerson, J. T., and Reichstein, M.: Global covariation of carbon turnover times with climate in terrestrial ecosystems, Nature, 514, 213-217, 10.1038/nature13731 2014.

Chen, Y., Morton, D. C., Andela, N., van der Werf, G. R., Giglio, L., and Randerson, J. T.: A pan-tropical cascade of fire driven by El Niño/Southern Oscillation, Nature Climate Change, 7, 906-911, 10.1038/s41558-017-0014-8, 2017.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) system: design, theory, and implementation, Journal of Advances in Modeling Earth Systems, 10, 2731-2754, 2018.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., and Krinner, G.: Long-term climate change: projections, commitments and irreversibility, in: Climate Change 2013-The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 1029-1136, 2013.

Chuvieco, E., Lizundia-Loiola, J., Pettinari, M. L., Ramo, R., Padilla, M., Tansey, K., Mouillot, F., Laurent, P., Storm, T., Heil, A., and Plummer, S.: Generation and analysis of a new global burned area product based on MODIS 250&thinsp;m reflectance bands and thermal anomalies, Earth Syst. Sci. Data, 10, 2015-2031, 10.5194/essd-10-2015-2018, 2018.

Claverie, M., Matthews, J. L., Vermote, E. F., and Justice, C. O.: A 30+ Year AVHRR LAI and FAPAR Climate Data Record: Algorithm Description and Validation, Remote Sensing, 8, 263, 2016.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, 10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Harris, N. R. P., Rex, M., Shepherd, T. G., Fahey, D. W., Amanatidis, G. T., Austin, J., Chipperfield, M. P., Dameris, M., Forster, P. M. D. F., Gettelman, A., Graf, H. F., Nagashima, T., Newman, P. A., Pawson, S., Prather, M. J., Pyle, J. A., Salawitch, R. J., Santer, B. D., and Waugh, D. W.: A Strategy for Process-Oriented Validation of Coupled Chemistry–Climate Models, Bulletin of the American Meteorological Society, 86, 1117-1134, 10.1175/bams-86-8-1117, 2005.

Forkel, M., Andela, N., Harrison, S. P., Lasslop, G., van Marle, M., Chuvieco, E., Dorigo, W., Forrest, M., Hantson, S., Heil, A., Li, F., Melton, J., Sitch, S., Yue, C., and Arneth, A.: Emergent relationships with respect to burned area in global satellite observations and fire-enabled vegetation models, Biogeosciences, 16, 57-76, 10.5194/bg-16-57-2019, 2019.

Forkel, M., Dorigo, W., Lasslop, G., Chuvieco, E., Hantson, S., Heil, A., Teubner, I., Thonicke, K., and Harrison, S. P.: Recent global and regional trends in burned area and their compensating environmental controls, Environmental Research Communications, 1, 051005, 10.1088/2515-7620/ab25d2, 2019.

French, N. H. F., de Groot, W. J., Jenkins, L. K., Rogers, B. M., Alvarado, E., Amiro, B., de Jong, B., Goetz, S., Hoy, E., Hyer, E., Keane, R., Law, B. E., McKenzie, D., McNulty, S. G., Ottmar, R., Pérez-Salicrup, D.

R., Randerson, J., Robertson, K. M., and Turetsky, M.: Model comparisons for estimating carbon emissions from North American wildland fire, Journal of Geophysical Research: Biogeosciences, 116, G00K05, 10.1029/2010jg001469, 2011.

Friend, A. D., Lucht, W., Rademacher, T. T., Keribin, R., Betts, R., Cadule, P., Ciais, P., Clark, D. B., Dankers, R., Falloon, P. D., Ito, A., Kahana, R., Kleidon, A., Lomas, M. R., Nishina, K., Ostberg, S., Pavlick, R., Peylin, P., Schaphoff, S., Vuichard, N., Warszawski, L., Wiltshire, A., and Woodward, F. I.: Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO2, Proceedings of the National Academy of Sciences, 111, 3280-3285, 10.1073/pnas.1222477110, 2014.

Giglio, L., Randerson, J. T., and Werf, G. R.: Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4), Journal of Geophysical Research: Biogeosciences, 118, 317-328, 2013.

Giorgi, F., and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the “Reliability Ensemble Averaging” (REA) Method, Journal of Climate, 15, 1141-1158, 10.1175/1520-0442, 2002.

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.

Hantson, S., Pueyo, S., and Chuvieco, E.: Global fire size distribution is driven by human impact and climate, Global Ecology and Biogeography, 24, 77-86, 10.1111/geb.12246, 2015.

Hantson, S., Arneth, A., Harrison, S. P., Kelley, D. I., Prentice, I. C., Rabin, S. S., Archibald, S., Mouillot, F., Arnold, S. R., Artaxo, P., Bachelet, D., Ciais, P., Forrest, M., Friedlingstein, P., Hickler, T., Kaplan, J. O., Kloster, S., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Melton, J. R., Meyn, A., Sitch, S., Spessa, A., van der Werf, G. R., Voulgarakis, A., and Yue, C.: The status and challenge of global fire modelling, Biogeosciences, 13, 3359-3375, 10.5194/bg-13-3359-2016, 2016.

Held, I. M.: The Gap between Simulation and Understanding in Climate Modeling, Bulletin of the American Meteorological Society, 86, 1609-1614, 10.1175/bams-86-11-1609, 2005.

Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A., Djalante, R., Ebi, K., and Engelbrecht, F.: Impacts of 1.5 ºC global warming on natural and human systems, In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, W.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, W. Chen, X. Zhou, M.I.Gomis, E. Lonnoy, T.Maycock, M.Tignor, and T. Waterfield (eds.)], 2019.

Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlström, A., Arneth, A., Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Jain, A. K., Kato, E., Papale, D., Poulter, B., Raduly, B., Rödenbeck, C., Tramontana, G., Viovy, N., Wang, Y.-P., Weber, U., Zaehle, S., and Zeng, N.: Compensatory water effects link yearly global land CO2 sink changes to temperature, Nature, 541, 516-520, 10.1038/nature20780, 2017.

Kaiser, J. W., Heil, A., Andreae, M. O., Benedetti, A., Chubarova, N., Jones, L., Morcrette, J. J., Razinger, M., Schultz, M. G., Suttie, M., and van der Werf, G. R.: Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power, Biogeosciences, 9, 527-554, 10.5194/bg-9-527-2012, 2012.

Kelley, D. I., Bistinas, I., Whitley, R., Burton, C., Marthews, T. R. and Dong, N.: How contemporary bioclimatic and human controls change global fire regimes, Nat. Clim. Chang., 9(9), 690–696, 2019.

Kelley, D., Prentice, I. C., Harrison, S., Wang, H., Simard, M., Fisher, J., and Willis, K.: A comprehensive benchmarking system for evaluating global vegetation models, Biogeosciences, 10, 3313-3340, 2013.

Kirtman, B., Power, S., Adedoyin, A., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., and Meehl, G.: Near-term climate change: projections and predictability, In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, W. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Kloster, S., and Lasslop, G.: Historical and future fire occurrence (1850 to 2100) simulated in CMIP5 Earth System Models, Global and Planetary Change, 150, 58-69, http://doi.org/10.1016/j.gloplacha.2016.12.017, 2017.

Knorr, W., Jiang, L., and Arneth, A.: Climate, CO2 and human population impacts on global wildfire emissions, Biogeosciences, 13, 267-282, 10.5194/bg-13-267-2016, 2016.

Knutti, R.: The end of model democracy?, Climatic Change, 102, 395-404, 10.1007/s10584-010-9800-2, 2010.

Lasslop, G., Coppola, A. I., Voulgarakis, A., Yue, C., and Veraverbeke, S.: Influence of Fire on the Carbon Cycle and Climate, Current Climate Change Reports, 5, 112-123, 10.1007/s40641-019-00128-9, 2019.

Lasslop, G., Thonicke, K., and Kloster, S.: SPITFIRE within the MPI Earth system model: Model development and evaluation, Journal of Advances in Modeling Earth Systems, 6, 740-755, 10.1002/2013ms000284, 2014.

Le Page, Y., Oom, D., Silva, J. M. N., Jönsson, P., and Pereira, J. M. C.: Seasonality of vegetation fires as modified by human action: observing the deviation from eco-climatic fire regimes, Global Ecology and Biogeography, 19, 575-588, 10.1111/j.1466-8238.2010.00525.x, 2010.

Le Page, Y., Morton, D., Bond-Lamberty, B., Pereira, J. M. C., and Hurtt, G.: HESFIRE: a global fire model to explore the role of anthropogenic and weather drivers, Biogeosciences, 12, 887-903, 10.5194/bg-12-887-2015, 2015.

Lehsten, V., Tansey, K., Balzter, H., Thonicke, K., Spessa, A., Weber, U., Smith, B., and Arneth, A.: Estimating carbon emissions from African wildfires, Biogeosciences, 6, 349-360, 10.5194/bg-6-349-2009, 2009.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367-371, 10.1038/nature15371, 2015.

Li, F., Levis, S., and Ward, D.: Quantifying the role of fire in the Earth system–Part 1: Improved global fire modeling in the Community Earth System Model (CESM1), Biogeosciences, 10, 2293-2314, 2013.

Li, F., Val Martin, M., Hantson, S., Andreae, M. O., Arneth, A., Lasslop, G., Yue, C., Bachelet, D., Forrest, M., Kaiser, J. W., Kluzek, E., Liu, X., Melton, J. R., Ward, D. S., Darmenov, A., Hickler, T., Ichoku, C., Magi, B. I., Sitch, S., van der Werf, G. R., and Wiedinmyer, C.: Historical (1700-2012) Global Multi-model Estimates of the Fire Emissions from the Fire Modeling Intercomparison Project (FireMIP), Atmos. Chem. Phys. Discuss., 2019, 1-57, 10.5194/acp-2019-37, 2019.

Luyssaert, S., Inglima, I., Jung, M., Richardson, A. D., Reichsteins, M., Papale, D., Piao, S. L., Schulzes, E. D., Wingate, L., Matteucci, G., Aragao, L., Aubinet, M., Beers, C., Bernhoffer, C., Black, K. G., Bonal, D., Bonnefond, J. M., Chambers, J., Ciais, P., Cook, B., Davis, K. J., Dolman, A. J., Gielen, B., Goulden, M., Grace, J., Granier, A., Grelle, A., Griffis, T., Gruenwald, T., Guidolotti, G., Hanson, P. J., Harding, R., Hollinger, D. Y., Hutyra, L. R., Kolar, P., Kruijt, B., Kutsch, W., Lagergren, F., Laurila, T., Law, B. E., Le Maire, G., Lindroth, A., Loustau, D., Malhi, Y., Mateus, J., Migliavacca, M., Misson, L., Montagnani, L., Moncrieff, J., Moors, E., Munger, J. W., Nikinmaa, E., Ollinger, S. V., Pita, G., Rebmann, C., Roupsard, O., Saigusa, N., Sanz, M. J., Seufert, G., Sierra, C., Smith, M. L., Tang, J., Valentini, R., Vesala, T., and Janssens, I. A.: CO(2) balance of boreal, temperate, and tropical forests derived from a global database, Global Change Biology, 13, 2509-2537, 10.1111/j.1365-2486.2007.01439.x, 2007.

Magi, B. I., Rabin, S., Shevliakova, E., and Pacala, S.: Separating agricultural and non-agricultural fire seasonality at regional scales, Biogeosciences, 9, 3003-3012, 10.5194/bg-9-3003-2012, 2012.

Mangeon, S., Voulgarakis, A., Gilham, R., Harper, A., Sitch, S., and Folberth, G.: INFERNO: a fire and emissions scheme for the Met Office's Unified Model, Geosci. Model Dev. Discuss., 2016, 1-21, 10.5194/gmd-2016-32, 2016.

Melton, J. R., and Arora, V. K.: Competition between plant functional types in the Canadian Terrestrial Ecosystem Model (CTEM) v. 2.0, Geosci. Model Dev., 9, 323-361, 10.5194/gmd-9-323-2016, 2016.

Michaletz, S. T., Cheng, D., Kerkhoff, A. J., and Enquist, B. J.: Convergence of terrestrial plant production across global climate gradients, Nature, 512, 39-43, 10.1038/nature13470, 2014.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani, R. R., and Running, S. W.: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sensing of Environment, 83, 214-231, https://doi.org/10.1016/S0034-4257(02)00074-3, 2002.

Olson, R., Johnson, K., Zheng, D., and Scurlock, J.: Global and regional ecosystem modeling: Databases of model drivers and validation measurements, ORNL Technical Memorandum TM-2001/196. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, 2001.

Parker, W. S.: Ensemble modeling, uncertainty and robust predictions, Wiley Interdisciplinary Reviews: Climate Change, 4, 213-223, 10.1002/wcc.220, 2013.

Pettinari, M. L., and Chuvieco, E.: Generation of a global fuel data set using the Fuel Characteristic Classification System, Biogeosciences, 13, 2061-2076, 10.5194/bg-13-2061-2016, 2016.

Pfeiffer, M., Spessa, A., and Kaplan, J.: A model for global biomass burning in preindustrial time: LPJ-LMfire (v1. 0), Geoscientific Model Development, 6, 643-685, 2013.

Rabin, S. S., Melton, J. R., Lasslop, G., Bachelet, D., Forrest, M., Hantson, S., Kaplan, J. O., Li, F., Mangeon, S., Ward, D. S., Yue, C., Arora, V. K., Hickler, T., Kloster, S., Knorr, W., Nieradzik, L., Spessa, A., Folberth, G. A., Sheehan, T., Voulgarakis, A., Kelley, D. I., Prentice, I. C., Sitch, S., Harrison, S., and Arneth, A.: The Fire Modeling Intercomparison Project (FireMIP), phase 1: experimental and analytical protocols with detailed model descriptions, Geosci. Model Dev., 10, 1175-1197, 10.5194/gmd-10-1175-2017, 2017.

Randerson, J., Chen, Y., Werf, G., Rogers, B., and Morton, D.: Global burned area and biomass burning emissions from small fires, Journal of Geophysical Research: Biogeosciences (2005–2012), 117, 2012.

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.

Roy, D. P., Boschetti, L., Justice, C. O., and Ju, J.: The collection 5 MODIS burned area product — Global evaluation by comparison with the MODIS active fire product, Remote sensing of environment, 112, 3690-3707, 2008.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and Morel, A.: Benchmark map of forest carbon stocks in tropical regions across three continents, Proceedings of the National Academy of Sciences, 108, 9899-9904, 10.1073/pnas.1019576108, 2011.

Santer, B. D., Taylor, K. E., Gleckler, P. J., Bonfils, C., Barnett, T. P., Pierce, D. W., Wigley, T. M. L., Mears, C., Wentz, F. J., Brüggemann, W., Gillett, N. P., Klein, S. A., Solomon, S., Stott, P. A., and Wehner, M. F.: Incorporating model quality information in climate change detection and attribution studies, Proceedings of the National Academy of Sciences, 106, 14778-14783, 10.1073/pnas.0901736106, 2009.

Settele, J., Scholes, R., Betts, R., Bunn, S., Leadley, P., Nepstad, D., Overpeck, J. T., and Taboada, M. A.: Terrestrial and inland water systems, in: Climate change 2014: Impacts,adaptation, and vulnerability. Part a: Global and sectoral aspects. Contribution of working group ii to the fifth assessment report of the intergovernmental panel on climate change, edited by: Field, C. B., Barros, V. R., Dokken, D. J., Mach, K. J., Mastrandrea, M. D., Bilir, T. E., Chatterjee, M., Ebi, K. L., Estrada, W. O., Genova, R. C., Girma, B., Kissel, E. S., Levy, A. N., MacCracken, S., Mastrandrea, P. R., and L.L., W., Cambridge University Press, Cambridge, 271-359, 2014.

Smith, B., Wårlind, D., Arneth, A., Hickler, T., Leadley, P., Siltberg, J., and Zaehle, S.: Implications of incorporating N cycling and N limitations on primary production in an individual-based dynamic vegetation model, Biogeosciences, 11, 2027-2054, 10.5194/bg-11-2027-2014, 2014.

Song, X., Li, F., Harrison, S.P., Arneth, A., Forrest, M., Hantson, S., Lasslop, G., Mangeon, S., Melton, J.R., Yue, C., Hickler, T., Xu, X. Assessment of vegetation biomass and its 20th century changes over China based on a combination of multi-model simulations and field observations. *Environmental Research Letters*, subm., 2019.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bull. Am. Meteorol. Soc., 93, DOI:0.1175/BAMS-D-1111-00094.00091, 2012.

Teckentrup, L., Harrison, S. P., Hantson, S., Heil, A., Melton, J. R., Forrest, M., Li, F., Yue, C., Arneth, A., Hickler, T., Sitch, S., and Lasslop, G.: Sensitivity of simulated historical burned area to environmental andanthropogenic controls: A comparison of seven fire models, Biogeosciences Discuss., 1-39, 10.5194/bg-2019-42, accepted, 2019.

Thurner, M., Beer, C., Santoro, M., Carvalhais, N., Wutzler, T., Schepaschenko, D., Shvidenko, A., Kompter, E., Ahrens, B., Levick, S. R., and Schmullius, C.: Carbon stock and density of northern boreal and temperate forests, Global Ecology and Biogeography, 23, 297-310, 10.1111/geb.12125, 2014.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291-4313, 10.5194/bg-13-4291-2016, 2016.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Mu, M., Kasibhatla, P. S., Morton, D. C., DeFries, R. S., Jin, Y., and van Leeuwen, T. T.: Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997-2009), Atmospheric Chemistry and Physics, 10, 11707-11735, DOI 10.5194/acp-10-11707-2010, 2010.

van Leeuwen, T. T., van der Werf, G. R., Hoffmann, A. A., Detmers, R. G., Rücker, G., French, N. H. F., Archibald, S., Carvalho Jr, J. A., Cook, G. D., de Groot, W. J., Hély, C., Kasischke, E. S., Kloster, S., McCarty, J. L., Pettinari, M. L., Savadogo, P., Alvarado, E. C., Boschetti, L., Manuri, S., Meyer, C. P., Siegert, F., Trollope, L. A., and Trollope, W. S. W.: Biomass burning fuel consumption rates: a field measurement database, Biogeosciences, 11, 7305-7329, 10.5194/bg-11-7305-2014, 2014.

Warszawski, L., Friend, A., Ostberg, S., Frieler, K., Lucht, W., Schaphoff, S., Beerling, D., Cadule, P., Ciais, P., Clark, D. B., Kahana, R., Ito, A., Keribin, R., Kleidon, A., Lomas, M., Nishina, K., Pavlick, R., Rademacher, T. T., Buechner, M., Piontek, F., Schewe, J., Serdeczny, O., and Schellnhuber, H. J.: A multi-model analysis of risk of ecosystem shifts under climate change, Environmental Research Letters, 8, 044018, 10.1088/1748-9326/8/4/044018, 2013.

Ward, D. S., Shevliakova, E., Malyshev, S., Lamarque, J. F., and Wittenberg, A. T.: Variability of fire emissions on interannual to multi-decadal timescales in two Earth System models, Environmental Research Letters, 11, 125008, 2016.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, Q J Roy Meteor Soc, 134, 241-260, 10.1002/qj.210, 2008.

Wu, M., Knorr, W., Thonicke, K., Schurgers, G., Camia, A., and Arneth, A.: Sensitivity of burned area in Europe to climate change, atmospheric CO2 levels, and demography: A comparison of two fire-vegetation models, Journal of Geophysical Research: Biogeosciences, 120, 2256-2272, 2015.

Yue, C., Ciais, P., Cadule, P., Thonicke, K., Archibald, S., Poulter, B., Hao, W. M., Hantson, S., Mouillot, F., Friedlingstein, P., Maignan, F., and Viovy, N.: Modelling the role of fires in the terrestrial carbon balance by incorporating SPITFIRE into the global vegetation model ORCHIDEE – Part 1: simulating historical global burned area and fire regimes, Geosci. Model Dev., 7, 2747-2767, 10.5194/gmd-7-2747-2014, 2014.

Table 1: Brief description of the global fire models that ran the FireMIP baseline experiments. Process indicates models which explicitly simulate ignitions and fire spread. A detailed overview can be found in Rabin et al. (2017).

Model	*Dynamic Biogeography*	*Fire model type*	*Human suppression of fire spread/ burnt area*	*Spatial resolution (lon x lat)*	*Temporal resolution*	*Reference*
CLM	No	Process	Yes	2.5° x 1.9°	Half hourly	Li et al., 2013
CLASS-CTEM	No	Process	Yes	2.8125° x 2.8125°	Daily	Melton and Arora, 2016
JULES-INFERNO	Yes, but without fire feedback	Empirical	No	1.875° x 1.245°	Half hourly	Mangeon et al., 2016
JSBACH-SPITFIRE	No	Process	Yes	1.875° x 1.875°	Daily	Lasslop et al., 2014
LPJ-GUESS-SPITFIRE	Yes	Process	No	0.5° x 0.5°	Daily	Lehsten et al., 2009
LPJ- GUESS-GlobFIRM	Yes	Empirical	No	0.5° x 0.5°	Annual	Smith et al., 2014
LPJ-GUESS-SIMFIRE-BLAZE	Yes	Empirical	Yes	0.5° x 0.5°	Annual	Knorr et al., 2016
MC2	Yes	Process	No	0.5° x 0.5°	Monthly	Bachelet et al., 2015
ORCHIDEE-SPITFIRE	No	Process	No	0.5° x 0.5°	Daily	Yue et al., 2014

Table 2: Simulated and observed burnt area (Mha) for the period 2002-2012 for the globe and for key regions including the northern extratropics (NET, > 30°N), the southern extra (SET, > 30°S), the tropics (30°N - 30°S), the savanna regions of Africa (18°W-40°E & 13°N-20°S), the savanna region of South America (42°-68°W & 9°S-25°S), Australian Savanna 155°E & 11°S-20°S), the agricultural band of central Eurasia (30°E-85°E & 50°N-58°N), the Mediterranean basin (10°W-37°E & 31°N-44°N), and the western USA (100°-125°W & 43°N). Data availability for FireCCI40 is limited to 2005-2011 and for MC2 to 2002-2009.

	Global	NET	Tropics	SET	S-American savanna	African savanna	Australian savanna	Central Eurasia	Mediterranean basin	wes
GFED4s	468	39	426	4	18	295	35	8.5	1.3	
GFED4	349	27	319	3	14	218	34	5.2	0.8	
MCD45	348	33	312	4	13	232	25	7.0	2.0	
FireCCI40	345	23	320	2	8	237	25	6.8	1.1	
FireCCI51	387	37	347	3	14	230	38	10.2	1.3	
CLM	454	77	362	15	36	194	15	7.9	9.3	
CLASS-CTEM	536	41	466	28	46	172	20	2.0	4.3	
JULES-INFERNO	381	76	292	13	26	128	23	5.0	11.0	
JSBACH-SPITFIRE	457	114	318	25	21	166	17	15.5	9.5	
LPJ-GUESS-GlobFIRM	39	14	24	1	3	7	3	0.6	0.6	
LPJ-GUESS-SPITFIRE	393	99	280	14	51	135	2.8	12.5	14.5	
LPJ-GUESS-SIMFIRE-BLAZE	482	86	381	15	72	146	27	3.4	7.9	
MC2	97	40	54	3	2	17	2	0.9	5.0	
ORCHIDEE-SPITFIRE	471	16	435	19	13	246	81	2.4	2.4	

Table 3: Benchmarking scores after removing the influence of differences in the mean and variance for each individual global fire model for key fire and vegetation variables. A lower score is "better", with a perfect score equal to 0. The full table with all benchmarking scores is presented in Table S1. Dataset information can be found in Supplementary Information S1. LPJ-G = LPJ-GUESS. Cell are coloured blue if the benchmarking score is lower than both null models, yellow if lower than 1 null model and red when higher than both null models.

	Dataset	*Mean*	*random*	*CLM*	*CLASS-CTEM*	*JULES INFERNO*	*JSBACH SPITFIRE*	*LPJ-G GlobFIRM*	*LPJ-G SPITFIRE*	*SIMFIRE-BLAZE*	*MC2*	*ORCHIDEE SPITFIRE*
Burnt area												
spatial	GFED4s	1	1.07	0.63	0.79	0.72	0.70	1.06	0.94	0.88	1.00	0.72
	GFED4	1	1.14	0.80	0.93	0.85	0.86	1.08	0.98	0.88	1.07	0.71
	MCD45	1	1.16	0.65	0.81	0.72	0.69	1.12	0.93	0.92	1.02	0.70
	FireCCI40	1	1.13	0.77	0.98	0.89	0.92	1.09	0.93	0.97	1.13	0.73
	FireCCI51	1	1.11	0.83	1.01	0.91	0.93	1.11	0.96	0.97	1.23	0.70
seasonal phase	GFED4s	0.56	0.22	0.12	0.12	0.13	0.12		0.31			0.31
	GFED4	0.49	0.47	0.34	0.35	0.41	0.42		0.33			0.31
	MCD45	0.56	0.26	0.12	0.11	0.12	0.12		0.30			0.30
	FireCCI40	0.60	0.12	0.16	0.43	0.17	0.16		0.33			0.32
	FireCCI51	0.55	0.25	0.26	0.28	0.33	0.32		0.32			0.31
seasonal concentration	GFED4s	1	1.36	1.16	1.15	1.24	1.15		1.13			1.22
	GFED4	1	1.35	1.19	1.12	1.25	1.11		1.18			1.19
	MCD45	1	1.36	1.14	1.08	1.26	1.13		1.12			1.20
	FireCCI40	1	1.34	1.31	1.26	1.36	1.25		1.29			1.30
	FireCCI51	1	1.36	1.25	1.22	1.33	1.21		1.20			1.27
IAV	GFED4s	1	1.46	1.17	0.65	1.18	1.09	0.66	1.36	0.76	1.66	1.44
	GFED4	1	1.27	0.98	1.62	1.23	0.89	1.04	1.08	1.00	1.41	1.25
	MCD45	1	1.32	0.93	1.34	1.11	0.84	0.73	0.97	1.27	1.67	1.22
	FireCCI5.1	1	1.42	1.18	1.53	1.24	1.27	1.73	1.27	1.23	1.87	1.12
fire emission												
spatial	GFAS	1	1.08	0.78	0.85	0.73	0.74	1.13	1.03	0.91	1.06	0.86
seasonal phase	GFAS	0.78	0.18	0.16	0.20	0.17	0.15		0.37			0.34
seasonal concentration	GFAS	1	1.36	1.20	1.22	1.30	1.17		1.27			1.25
IAV	GFAS	1	1.36	0.77	1.70	1.28	1.09	1.42	1.42	1.11	1.41	1.49
Fire number												
spatial	Hantson	1	1.19				0.96		0.83			0.76
Fire size												
Spatial	Hantson	1	1.31				1.03		1.22			1.12
GPP												
spatial	Luyssaert	1	1.39	1.49	1.41	1.46	1.39	1.41	1.24	1.37		1.09
spatial	Jung	1	1.30	0.64	0.46	0.39	0.42	0.46	0.67	0.43		0.49
seasonal phase	Jung	0.42	0.65	0.18	0.23	0.19	0.23		0.22			0.22
seasonal concentration	Jung	1	1.65	1.08	1.19	1.14	1.21		1.19			1.09
NPP												
spatial	Michaletz	1	1.39	0.82	0.79	0.77	0.75	0.96	0.86	0.89	0.88	0.99
spatial	Luyssaert	1	1.33	0.90	1.01	0.53	0.76	0.82	0.87	0.79	0.68	0.84
spatial	EMDI	1	1.30	0.91	0.87	0.58	0.66	0.79	0.83	0.81	0.65	0.80
LAI												
spatial	MCD15	1	1.29	0.60	0.53	0.44	0.78	0.70	0.61	0.57	0.63	
spatial	AVHRR	1	1.29	0.81	0.71	0.49	0.65	0.74	0.62	0.61	0.64	
Carbon in vegetation												
spatial	Avitabile	1	1.32	0.69	0.88	0.76	0.78	0.76	0.76	0.74	0.80	0.70
spatial	Carvalhais	1	1.32	0.66	0.66	0.58	0.64	0.62	0.66	0.58	0.67	0.54

Formatted Table

Formatted: Pattern: Clear (Custom Color(RGB(223,238,246)))