Dear editor and referees,

We want to thank you for your thoughts and comments on this manuscript. The reviews helped to clarify and improve the methodology, and reflect on the novel conclusions from this study compared to previous findings.

The major changes to the manuscript therefore are:

- A better explanation of the scope and novelty of this study (in the introduction, the discussion and conclusion sections)
- A clarification of the analysis of the trend in burned area and improved consistency between the different forcing factors

We below address the reviewer's comments point by point. We add *our replies in italic* and highlight suggested modifications in the manuscript in red. We number our replies and cross-refer to them to reduce the text if points had already been addressed before.

Referee #2

General comments

The study is a useful compilation of the analysis of sensitivity experiments in the FireMIP output, but it is largely a technical report of the sensitivity of FireMIP model simulations of burned area since 1900. Philosophically, there is nothing really offered by the authors in terms of specific testing of improvements/changes needed with firemodels beyond what has been pointed out in the literature in papers such as Van Marle et al 2017 and Andela et al 2017, and hinted at in the Hantson et al 2016 FireMIP overview paper and the Forkel et al 2019 paper. While I appreciate the depth of the dissection of the causes for the discrepancies among FireMIP models in this study, I find myself with no questions about FireMIP that have new or interesting answers, which is a concerning lack of momentum from the initially promising FireMIP effort. For example, did the FireMIP sensitivity experiments produce knowledge that the modeling groups could leverage for specific technical advances on, say, a future set of experiments? If anything, this paper makes me increasingly skeptical about the utility of FireMIP other than to show precisely what these authors stated in their conclusions: "Although burned area in most models compares reasonably well with satellite observations, there is a huge spread in transient simulations before the satellite era and a huge spread in the influence of the driving factors between models." Again, however, many FireMIP related papers have already pointed this out.I recommend that the paper be published and I think that my comments fall somewhere between a minor and major revision, so I labeled it as

minor revisions even though some of my comments might require some major discussion amongst the authors in terms of structuring a reply or rebuttal. The challenge that I offer to the authors is this: I do not see what we gain beyond now knowing that the sensitivity experiments areas confusingly inconclusive as the core experiments. If I were re-formulating my firemodel and looking to this study, I would have little idea as to what the focus point should be other than simply acknowledging weaknesses such as the representation of human use of fire or needed better data for model parameterizations. The authors may need to make their case more clearly for this paper to stand out beyond being a technical report out.

1) We thank the reviewer for the critical review and take the chance to reflect and rework our conclusions. We include improvements in the Introduction, the discussion and the conclusions to clarify the novelty of our study.

In the introduction we clarify how our work relates to previous work:

Fire-enabled vegetation models simulate fire regimes in response to the combination of individual forcings, including atmospheric CO2 concentration, population density, land-use change, lightning and climate.

Individual fire-enabled vegetation models have been shown to simulate observed global patterns of burned area and fire emissions reasonably well (Kloster et al., 2010; Prentice et al., 2011; Li et al., 2012; Lasslop et al., 2014; Yue et al., 2014), but there are large differences between models in terms of regional patterns, fire seasonality and interannual variability, and historical trends (Kelley et al., 2013; Andela et al., 2017) and responses to individual factors (Kloster et al., 2010; Knorr et al., 2014, 2016; Lasslop and Kloster, 2017, 2015). The fire model intercomparison project (FireMIP, Hantson et al., 2016a; Rabin et al., 2017a) provides a systematic framework to consistently analyse and understand the causes of these differences and to relate them to differences in the treatment of key drivers of fire in individual models. The FireMIP project provides simulations for a systematic comparison of fire-model behaviour based on outputs of a large range of models with identical forcing inputs. In addition to a reference historical simulation, sensitivity simulations were conducted for individual forcings, specifically atmospheric CO2 concentration, population density, land-use change, lightning and climate. A recent evaluation of the FireMIP models indicates that the relationship with climatic parameters is captured well by models, the response to human factors is captured by some models and the response to vegetation productivity or the allocation of carbon to fuels needs refinement for most models (Forkel et al., 2019a). Comparisons of the FireMIP historical simulations found differences in transient model behaviour in the 20th century (Andela et al., 2017; van Marle et al., 2017). The causes of the differences and the reasons why different models show different responses are not yet understood.

Our study shows in detail which model responses of burned area to environmental factors can be understood, how these are related to the model equations and how these translate into certain trends of burned area. The understanding on how certain model assumptions lead to trends in burned area is novel, the need for this was emphasized by the previous publications (but they do not provide it) and the recently detected trends in the satellite data. We improved the sections discussing the new possibilities for model reparameterization:

The main concern for model applications is the large spread of the historical simulated burned area. It remains difficult to evaluate and optimize the transient burned area simulations as the period observed by satellites is still short and the trends are not robust (Forkel et al., 2019b). Fire proxies (charcoal and ice-cores) give information on biomass burning over longer time scales. They do not confirm the recent decrease in burned area detected by satellites, but also only contain very few datapoints for that period (Marlon et al., 2016). For a valid comparison with the long term fire proxies, including estimates of deforestation fires in the models will be crucial, as land-use change fire emissions likely have a strong contribution to the signal (Marlon et al., 2008). An improved understanding of uncertainties in observed trends of fire regimes is therefore necessary. Only robust information should be included in models.

Our analysis shows which parts of the models are particularly important to simulate changes in burned area and need additional observational constraints or improved process understanding. In line with previous research (Bistinas et al., 2014; Hantson et al., 2016a, b; Andela et al., 2017), the large divergence in the response to human activities between the FireMIP models shows that the human impact on fires is still insufficiently understood and therefore not constrained in current models.

specifically for the effect of land-use change on burned area:

We identify land-use change as the major cause of inter-model spread. Only one model explicitly includes fires associated with land-use and land cover change (cropland and deforestation fires), all the other models only include such effects through changes in vegetation parameters and structure. The inclusion of cropland fires is certainly important to understand and project changes in emissions, air pollution and the carbon cycle (Li et al., 2018; Arora and Melton, 2018). Cropland fires are, due to their small extent and low intensity, still a major uncertainty in our current understanding of global burned area (Randerson et al., 2012). Biases in the spatial patterns of burned area and the relationship between cropland fraction and burned area can therefore be expected. High resolution remote sensing may help to improve the detection (Hall et al., 2016). Moreover, understanding why and when humans

burn croplands on a regional scale may help to find an adequate representation of cropland fires within models and avoid overfitting to observational datasets. As croplands are simply excluded from burning in most models (except two), the spread of the other models is likely related to the treatment of pastures. Fires on pasturelands have been estimated to contribute over 40% of the global burned area (Rabin et al., 2015). Pasture fires are not treated explicitly in any of the models, although some models slightly modify the vegetation on pastures by harvesting or changing the fuel bulk density (see tab. 5). Expansion of pastures is mostly implemented by simply increasing the area of grasslands. Information on how fuel properties differ between pastures and natural grasslands could therefore help to improve model parametrisations. Prescribing fires on anthropogenic land covers can be a solution for certain applications of fire models (Rabin et al., 2018). Grazing intensity was found to be related to decreases in burned area (Andela et al., 2017). Models so far represent the area that is converted due to land cover change but not the intensity of land-use. This was partly due to the lack of global data regarding land use intensity which is now becoming available and provides new opportunities for fire model development (e.g. the LUH2 dataset; Hurtt et al., 2017). In the sensitivity simulations shown here, even models that decrease burned area due to land-use and land cover change do not show a further decrease over the last decade. This indicates that model input datasets, explicit in time and space, for land-use intensity and grazing intensity are necessary for fire projections. The level of socioeconomic development also modifies the relationship between humans and burned area (Andela et al., 2017; Forkel et al., 2017). Regional analysis of remote sensing data could be highly useful, as a global relationship between burned area and individual human factors as assumed in many models and also statistical analysis is not likely. Assumptions on how different human groups (hunter-gatherers, pastoralists, and farmers) use fire have been included in a paleofire model (Pfeiffer et al., 2013). The development of such an approach for modern times would be highly valuable for fire models that aim to model the recent decades and future.

for the effect of CO2 on burned area:
We show that, although all models show an overall increase in biomass as a consequence of increasing atmospheric CO2 concentration, models disagree about whether this results in an increase or decrease in burned area. The disagreement reflects the complex ways in which changes in atmospheric CO2 concentration influence vegetation properties, which results in different responses in different ecosystems. For LPJ-GUESS-SPITFIRE and JSBACH-SPITFIRE the CO2 fertilization effect considerably contributed to an increase in burned area. Such an effect is so far only supported for fuel limited

areas (Forkel et al., 2019b). The assumption that the influence of higher fuel load on burned area levels off for high fuel loads as used in other models could help to reduce this increase in burned area in regions with higher fuel load.

for the effect of climate and lightning on burned area in general:
Climate and lightning have a much lower effect on the trends than the other factors. While this study focuses on the trends, research on the short term variability and extreme events will be highly useful to investigate fire risks. The influence of climate and lightning on fire are therefore important research topics even if we find a comparably low influence on the long term trends. Moreover the trends in climate parameters may increase for the future and therefore the influence on burned area might increase.

and for the effect of lightning on burned area specifically:
But not only spatial patterns of lightning are important, the co-variation with climate as well as the temporal resolution of the input dataset determine the influence on burned area (Felsberg et al., 2018). Although we do not detect large signals in global burned area due to changes in lightning, lightning is known to be an important cause of ignitions regionally and is potentially involved in more complex interactions between fire, vegetation and climate, which can speed up the northward expansion of trees to the north in boreal regions (Veraverbeke et al., 2017). Thus, although our results suggest that the influence of increasing lightning is negligible at a global scale, it is a potentially important factor for process-based models that aim to model interactions between fire, vegetation and climate.

In addition, we point to datasets that can be used for model evaluation:
Recent advances in remote sensing products have high potential to support model development. However, remotely sensed burned area datasets alone are not a sufficient basis to evaluate fire models as many model structures can lead to reasonable burned area patterns. The emergence of longer records of burned area and the increasing availability of information on other aspects of the fire regime considerably improve opportunities to evaluate and improve our models. The FRY database (Laurent et al., 2018) and the global fire atlas (Andela et al., 2018), for example provide information on fire size, numbers of fire, rate of spread, and the characteristics of fire patches. These datasets will be useful to, for instance, separate effects of ignition and suppression. Rate of spread equations in global fire models are at present either very simple empirical representations tuned to improve burned area or based on laboratory experiments (Hantson et al., 2016). The mentioned datasets now offer the opportunity to derive parameters for rate of spread

equations at the spatial scales these models operate on. Fire size and rate of spread are important target variables besides burned area that can determine the impacts of fire. The effects on vegetation (combustion of biomass and tree mortality; Williams et al., 1999; Wooster et al., 2005) and on the atmosphere (Veira et al., 2016) are a function of fire intensity, which is also included in the FRY database (Laurent et al., 2018). A better evaluation of such parameters can enhance the usability of fire model simulations.
The specific model application has a strong influence on judging the validity of a model. Our analyses of the controls on the variability of fire suggest that human activities drive the long term (decadal to centennial) trajectories, while considering climate variability may be sufficient for short-term projections. Changes in the trends of the driving factors may change this balance. For instance, stronger changes in climate into the future may increase the relative importance of climate for long term fire projections in the future.

We change our Summary and conclusions to:
This comprehensive analysis of the influences of climate, lightning, CO2, population density and land-use and land cover change provides improved understanding of the relation between simulated historical trends in burned area and process representations in the models. It shows in detail which model responses of burned area to environmental factors can be understood, how these are related to the model equations, and how these translate into trends of burned area for the historical period.

Followed by the summary of insights for the individual factors. We add for the effect of population density:
It would be useful to develop an approach that represents local human-fire relationships, but this will likely remain a long term challenge and requires the synthesis of knowledge from various research fields.

We add for the effect of land use and land cover change:
Improved knowledge on the effects of land-use intensity on burned area and the development of appropriate forcing datasets could strongly support model development.

And end with:
The uncertainties in global fire models need to be taken into account in model applications, for instance if model simulations are to be used to support climate adaptation strategies. Model ensemble simulations can give indications of such uncertainties. Therefore the results of this study provide a basis to interpret uncertainties in global fire modelling studies. The spatial patterns of burned area and its drivers are already well explored and

understood. We here provide a summary of which model assumptions need additional constraints to efficiently reduce the uncertainty in temporal trends.

Specific comments

Figures in the Supplement – please make larger versions of the maps in figures a1-a8. Another improvement would be to include a continuous rather than binary scale of values of the correlation coefficient in a2-a8. Painting the world with binary correlation coefficients would mask areas of potential weak and strong linear correlation. The strength of this study is the technical report-out of FireMIP sensitivity studies, so by making figures a1-a8 so hard to read, the authors are undermining the very purpose of the work. Read another way, the community may gain more with more detail in the manuscript.

2) Figure a2-a8 are not correlations but the slope coefficients. It only shows significant changes to identify regions with weak relationships. We wanted to emphasize the spatial distribution of decreases and increases and therefore chose this color scale. We now provide the graphs with the more detailed color scale and larger versions of the maps, because, as the reviewer suggests, it will be useful for the community.

Page 6 line 16-17 – authors stated they used a square root transformation to reduce the skewness of the distribution, but it is unclear why. Please expand on both the reasons and what this transformation accomplishes. Perhaps a supplemental figure?

3) See also reply 10 for reviewer 1. The correlation coefficient is most useful for normally distributed variables. The burned area varies over several orders of magnitude and the skewed distribution gives the highest importance to values with very high burned area. We transformed the data to improve the applicability of the metric. We include now a figure illustrating the influence of individual data points to the correlation, showing that the outliers in the untransformed data have a really high contribution and determine the correlation (figure A9 in the Appendix). This is improved with the squareroot transformation and would be further improved using a log transformation, but that would mean that grid cells with 0 would be excluded. With the transformation the contribution is better distributed to all data points, it is therefore more useful for global modelling where a too strong focus on only grid cells with high burned area can be distracting.

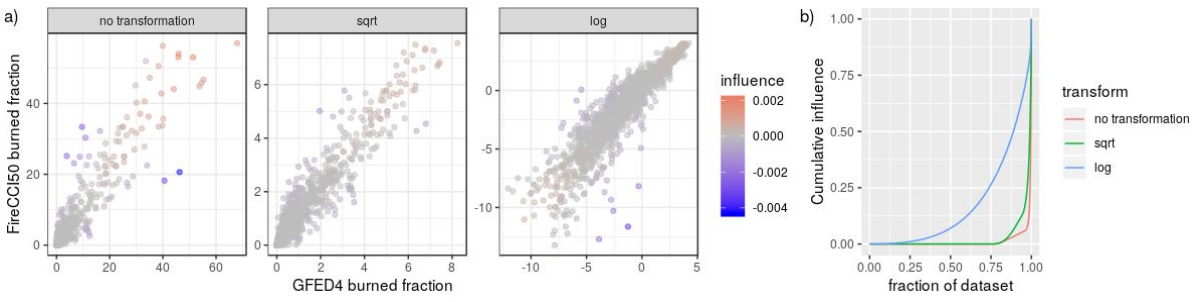

Figure A9: Scatter plots for the GFED4 and FireCCI50 dataset without transformation, square root transformation and log transformation (a), the color indicates the influence of individual data points on the correlation (computed as the difference in the correlation with and without that datapoint). Cumulative influence of data points in the dataset on the correlation (b). Without transformation a very small fraction has a strong influence on the correlation, these are grid cells with high burned area fraction (as can be seen in a).

We also modify the text in the main paper:
We quantify the agreement between models and observations by providing the global burned area and the Pearson correlation coefficient for the between grid cell variation (see tab. 3). We choose the Pearson correlation as it quantifies the covariation of the spatial patterns, and is less sensitive to the highly uncertain absolute burned area values. Burned area has a strongly skewed distribution, with few high values and many small values close to, or equal to, zero. These few high values have a much higher contribution to the overall correlation (see figure A9 in Appendix) and therefore the metric is strongly determined by the performance of the model in areas with high burning. Square root or logarithmic transformation leads to more normally distributed values, that reduce this bias (see figure A9 in Appendix). As the logarithm transformation excludes grid cells with zero burned area, we adopt the square root transformation.

Page 6 line 19 – major uncertainties is a subjective phrasing that requires more qualifications. Humber et al 2018 clearly discussed the nuanced and important ways that observed burned area data sets agree and disagree when using global, regional, and varying temporal scales. Looking at Figure 3 in Humber et al 2018 and Figure 1 in this paper, however, the implication is that FireMIP models have even more than “major” uncertainties in the sense that even at an annual time scale, there is more spread amongst models than amongst the observations. Furthermore, the three burned area data sets discussed in this study (GFED4, GFED4s, and FireCCI50) show that there is agreement unless the specific methodological approach is augmented with the small fires approach described in Randerson et al 2012. Is that really a major disagreement or just a difference in analysis? Please be more specific or careful in the discussion around observational uncertainties. Also, please see my comment about Figure 1 below.

4) See also reply 9 for reviewer 1. In Figure 1, the models are largely within the range of the observations for the evaluation period. The section shows that the models are largely in the range of satellite observed burned area and have a reasonable spatial distribution (see appendix figure A1). There is methodological uncertainty in satellite burned area products and this is reflected in the variation between the products due to the methodological

approach applied. The spread between these products still underestimates the uncertainty in the satellite products as all are based on the same sensor (MODIS). This is already mentioned in the manuscript on p.6 l. 23. We improve the paragraph with more details on the differences between the sensors and also link it to more recent burned area estimation using the high resolution Sentinel-2 data, which gives insights in the huge uncertainty of satellite products (see also reply 9 for reviewer 1).

To evaluate the simulations of burned area, we compare the simulated burned area with remote sensing data products. Global burned area observations from satellites still suffer from substantial uncertainty, as reflected by the considerable differences in spatial and temporal patterns between different data products (Humber et al., 2018; Hantson et al., 2016a; Chuvieco et al., 2018; van der Werf et al., 2017). Using multiple satellite products in model benchmarking is one approach to take into account these observational uncertainties (Rabin et al., 2017a). In this study, we use three satellite products: GFED4 (Giglio et al., 2013), GFED4s (van der Werf et al., 2017) and FireCCI50 (Chuvieco et al., 2018). GFED4 is a gridded version of the MODIS Collection 5.1 MCD64 burned area product. It is known that this product strongly underestimates small fires, including cropland fires (e.g.Hall et al. (2016)). In GFED4s, burned area due to small fires is estimated based on MODIS active fire (AF) detections and added to GFED4 burned area. However, this methodology may introduce significant errors related to erroneous AF detections (Zhang et al., 2018). As a complementary product, FireCCI50 was developed using MODIS spectral bands with higher spatial resolution than MCD64. A higher resolution enhances the ability to detect smaller fires; however, this improvement is partially offset by suboptimal spectral properties of the bands. Both GFED4s and FireCCI50 have larger burned area than GFED4. Since all three products are based on MODIS data, the inter-product differences probably underestimates uncertainties associated with these products. A recent mapping of burned area for Africa using higher resolution Sentinel-2 observations indicates that all three products substantially underestimate burned area (Roteta et al., 2019). For the model evaluation we use temporally averaged burned area fraction for the years 2001–2013, the interval common to all three satellite products and the model simulations.

Hall, J. V., T. V. Loboda, L. Giglio and G. W. McCarty (2016). "A MODIS-based burned area assessment for Russian croplands: Mapping requirements and challenges." Remote sensing of environment 184: 506-521.

Roteta, E., A. Bastarrika, M. Padilla, T. Storm and E. Chuvieco (2019). "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.

Zhang, T., Wooster, M., de Jong, M., and Xu, W.: How Well Does the 'Small Fire Boost' Methodology Used within the GFED4.1s Fire Emissions Database Represent the Timing, Location and Magnitude of Agricultural Burning?, Remote Sensing, 10, 823, https://doi.org/10.3390/rs10060823, 2018.

Moreover we now include a new publication (Forkel et al. 2019) in the discussion which shows that the trends as observed by satellites are still highly uncertain and not robust.
Satellite records show a decline in global burned area since 1996 (Andela et al., 2016). However, as Forkel et al. (2019b) have shown, the significance of the observed global decline is strongly affected by the length of the sampled interval because of the high interannual variability in burned area and trends between products show only a low correlation (Forkel et al., 2019b).
No observations document the longer term trends in burned area. Charcoal records (Marlon et al., 2008, 2016) and carbon monoxide data from ice-core records (Wang et al., 2010) are a proxy for biomass burning and show a global decrease in biomass burning over most of the 20th century. However, the charcoal records show an increase in burning since 2000 CE, but this discrepancy might reflect regional undersampling (for instance in Africa) or taphonomic issues of the charcoal record. A recent fire emission dataset (van Marle et al., 2017) merges information from satellites, charcoal records, airport visibility records and if no other information was available uses simulation results of the FireMIP models. This dataset is not included to evaluate the models here as it is partly based on the simulations of the FireMIP models and as it provides only estimates for emissions not burned area.
The understanding of the drivers on simulated trends that we give below provides insights on what causes the simulated trends and which assumptions control the trend. These insights will help to understand which observational constraints and process understanding is required to improve global fire models.

Page 6 line 20-21 – please explain what is meant by 0.01 and 0.2%. I am not following what the values refer to.

5) We clarify in the manuscript, see also reply 11 for reviewer 1:
[...] yields uncertainty estimates of 0.01 % (GFED4) and 0.2% (Fire CCI50)

Figure 1 would benefit from being split into a two-part plot: one part could remain asis, but the other would show the present day subset of the full analysis period. This is the evaluation period, but it is buried under too many curves.

6) Unfortunately this suggestion would lead to us exactly reproducing the figure number 3 of the Andela et al 2017 paper and contradicts the general suggestion of the reviewer to go beyond previous studies. We do agree, however, that the satellite datasets are buried under the curves in our plot. We now include a shaded area for the range of the satellite datasets as this is the main point we wish to convey here. As well, since we do not want to focus on evaluation of the models (which has been the focus of Andela et al. 2017 and Forkel et al. 2019 already) we rephrase the heading of this section to "Simulated historical burned area" to reflect the focus on the longer term trends and understanding the reasons for the divergence between models, independent of their correctness. We add a reference to Forkel et al. (2019) for more details.

Table 3 and page 7 – are these spatial correlation coefficients that compare the gridcell to grid cell agreement on a map? Or are they temporal correlation coefficients? It does not seem that Figure 1 temporal correlation is this high, but please clarify in the text. If this is a spatial correlation, please include the figure in the Appendix as it could be valuable to modelers in identifying regional weaknesses in the FireMIP simulated burned area.

7) We conduct a gridcell to gridcell comparison here, however spatial correlation coefficient is not a statistical term and may be confused with spatial auto-correlation. It implies some consideration of the geographical location. For table 3, we average burned area fraction over 2001 - 2013 (compare figure A1) and then correlate all individual grid cells of the remotely sensed product with the respective model. Therefore there is only one value, we did not analyse the spatial distribution or regional variation. For example, the first value in table 3, column 'R(GFED4, model)' is the Pearson correlation coefficient between the square root-transformed burned area fraction averaged over 2001 - 2013 in GFED4 and the square root-transformed burned area fraction averaged over 2001 - 2013 in CLASS-CTEM. We now include the "correlation over grid cells" to indicate it is not over time and change the caption of table 3 to "Global burned area averaged over 2001–2013 in Mha yr-1 and the Pearson correlation coefficients between burned area fraction averaged over 2001 - 2013 in the baseline experiment SF1 for all FireMIP-models and the respective observation data over all grid cells. We use a square root transformation on both model and observations. All correlation coefficients are significant (p-value < 0.05).

Table A2 is missing statistics relative to GFED4s.

8) GFED4s does not provide uncertainty estimates and therefore is not included in table A2. (We change the table caption from 'GFED4 and FireCCI50 provide uncertainty estimates' to 'Only GFED4 and FireCCI50

provide uncertainty estimates, therefore GFED4s is not included' to clarify this.)

Page 9 – the first sentence on this page highlights a major problem in the approach with modeling. Aiming at trends without a full understanding of the drivers in the simulations is .

9) One sentence in this comment is incomplete. It refers to the following sentence „The better understanding of the drivers of simulated trends that we provide below can inform us on how certain trends can be achieved in models." We speculate that the reviewer wants to indicate, that the possibility to achieve a trend based on a certain driver, does not necessarily mean that this is correct. Being aware however of how trends can be achieved is a useful information for model development. Whether the changes are plausible still needs to be addressed before implementing them.
We add:
The understanding of the drivers on simulated trends that we give below provides insights on what causes the simulated trends and which assumptions control the trend. These insights will help to understand which observational constraints and process understanding is required to improve global fire models.

Table 4 – while the M-K test is likely fine, the uncertainties (standard error or confidence intervals) in the slopes need to be included to understand the results better.

10) We include the uncertainties of the slope parameter. However the Mann-Kendall test is better suited to understand whether the trend is significant.

Page 9 and Section 3.2.4 – I thought that FireMIP only used a repeated lightning scaled to changes in modeled convection? While there is likely something to gain in the lightning sensitivity experiment, I would like to see some clearer discussion of the important caveats in interpreting the results. For example, would it be safe to surmise that there is no sensitivity to lightning changes since 1900 only if the modeled lightning is anything close to reality? Determining a lightning climatology from an untestable climate-model based parameterization and then drawing conclusions from that testing is prone to some circular or flawed logic.

11) The limitation of uncertainty in the lightning data is already included on p.20 line 10 where we see a major problem in conserving the correlation between lightning and other climate variables. We include now that the CAPE anomalies are derived from a global numerical weather prediction model. However, we don't see a flawed logic in showing that although the imposed lightning was strongly increasing the model results don't necessarily show

increases. That the present trend in the imposed lightning leads to a small change in burned area shows that the models have a low sensitivity to lightning. Lightning parameterizations of climate models are tested (see for instance Krause et al. (2014)). Krause et al. (2014) only show a decrease of lightning of 3.3% in pre-industrial times compared to present day. We add this information to give the reader an insight on the uncertainty. The results in Krause et al. (2014) however support our conclusion of the low sensitivity as they also only find small influences on burned area. Using the lightning dataset from Krause et al. (2014) instead of ours would likely reduce the response in burned area.

We add in the manuscript:

Most of the models show a low response of burned area to lightning (see fig. 2), although lightning rates increase by 20% over the simulation period - an increase that is much larger than the 3.3% change between pre-industrial times and the present estimated from a recent modelling study (Krause et al., 2014)

Figure 2 – please retitle these with something that is easier to quickly interpret without cross-referencing the table. For example, I suggest (a) Constant CO2 (SF2_CO2), (b) Constant Population (SF2_FPO), (c) Constant Land Cover (SF2_FLA), (d) Constant Lightning (SF2_FLI), (e) Constant climate (SF2_CLI). Also please make figure 2 much wider to avoid the visual clutter of overlaid zigzagging lines. & Figure 2 – change the y-axes ranges so they are constant. It is hard to understand the sensitivity if the plotted range is variable.

12) We changed the Figure according to the suggestions.

Page 11 line 9 – I agree that the statistics suggest individual trends are significant but this does not preclude the massive spread (both positive and negative) in the trends amongst models (table 4). I think this statement needs to include that caveat for an honest accounting of the FireMIP output.

13) The preceding sentence in the manuscript describes the details of the directions of the trends, including positive and negative trends.

Section 3.3 – the first paragraph makes no sense. What I am reading in this study is that the models barely agree on any trend, but yet the authors propose here that the models are important for understanding projected trends and supporting land management strategies. To me, a land management practice cannot be based on model trends that do not agree on trend and cannot be of much use if there is lack of agreement at country scales, let alone finer spatial scales.

14) We agree to some extent, that is why we wrote that the models need to be improved to be useful. We rephrase the paragraph and remove the reference to land management.

Global vegetation models are an important tool for examining the impacts of climate change and are used in policy-relevant contexts (IPCC, 2014; Schellnhuber et al., 2014; IPBES, 2016). Given the various influences of fire on the ecosystems (Bond et al., 2015), the carbon cycle and climate (Lasslop et al., 2019), improvements of global fire models are particularly important.

Section 3.3, second paragraph – the results presented in the manuscript clearly show that models only agree in magnitude in the present day, but the quick microscope analysis of the present day trends show that observations and models do not agree in trends. Some models predict a positive slope, some negative. Unless the authors intend to propose that one FireMIP model is more physically realistic than another, then the results of the sensitivity studies are inconclusive.

15) We agree with the reviewer that we cannot conclude from these analyses how the drivers caused real trends in fire regimes as the divergence between the models is too big. Only a few years ago it was not possible to detect any trends in the satellite data, the satellite estimate is still far from robust. The result of our sensitivity study is an improved understanding of how the trends are caused in the models and how certain trends can be achieved. We have rephrased the paragraph substantially, see reply 1.

Section 3.3 or 4 – it would be useful if these authors were to comment directly on fire models that did not contribute to FireMIP but that have contributed significantly to discussions of human-driven fire both in the present day and over the more distant past. This includes studies by Pfeiffer et al https://www.geosci-model-dev.net/6/643/2013/, Rabin et al https://www.geosci-model-dev.net/11/815/2018/, and Hantson et al https://journals.ametsoc.org/doi/full/10.1175/BAMS-D-15-00319.1 . All of these either echo or predict the results discussed by Andela et al 2017 and Bistinas et al 2014 related to a need to quantitatively represent the human use of fire on our planet in the modeling framework.

16) The previous papers acknowledged that the understanding of the human-fire relationship was rather low. However they could not provide the insight that this causes the largest divergence between global fire models as they were not based on a systematic comparison of simulation results. Moreover, we attribute specific model behaviour to the underlying model assumptions. We agree that some of these previous models give important information regarding incorporation of human-fire relationships (but Hantson et al. 2016 only summarizes the discussions of a workshop). Pfeiffer et al. (2013) deal with pre-industrial fire regimes. Rabin et al. (2018) is limited to the period of satellite observations, as they prescribe the agricultural burning based on satellite observations.

We integrate these earlier studies in section 3.3 and improve the discussion of the implications for model development. For the full context, see reply 1.

Our analysis shows which parts of the models are particularly important to simulate changes in burned area and need additional observational constraints or improved process understanding. In line with previous research (Bistinas et al., 2014; Hantson et al., 2016a, b; Andela et al., 2017), the large divergence in the response to human activities between the FireMIP models shows that the human impact on fires is still insufficiently understood and therefore not constrained in current models.

[...]

Fires on pasturelands have been estimated to contribute over 40% of the global burned area (Rabin et al., 2015). Pasture fires are not treated explicitly in any of the models, although some models slightly modify the vegetation on pastures by harvesting or changing the fuel bulk density (see tab. 5).

Expansion of pastures is mostly implemented by simply increasing the area of grasslands. Information on how fuel properties differ between pastures and natural grasslands could therefore help to improve model parametrisations.

Prescribing fires on anthropogenic land covers can be a solution for certain applications of fire models (Rabin et al., 2018).

[...]

Regional analysis of remote sensing data could be highly useful, as a global relationship between burned area and individual human factors as assumed in many models and also statistical analysis is not likely. Assumptions on how different human groups (hunter-gatherers, pastoralists, and farmers) use fire have been included in a paleofire model (Pfeiffer et al., 2013). The development of such an approach for modern times would be highly valuable for fire models that aim to model the recent decades and future.

[...]

Conclusions – the conclusions are already evident in the Andela et al 2017 paper,so I do not see what we gain in this study. The authors conclude “further analyses are required to better disentangle” factors, but this is the same conclusion so many firemodel and FireMIP papers have arrived at. Could the authors make a clearer argument about what we gain in this manuscript?

17) The cited phrase is not part of our conclusion sections, but part of the discussion. We delete it as it was not a substantial remark. For the gains of the manuscript see reply 1, 9, 16.