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 #1

The manuscript "Sensitivity of simulated historical burned area to environmental and anthropogenic controls: A comparison of seven fire models" by Teckentrup et al compares several global fire schemes implemented in different global land surface models in a controlled setup (based on FireMIP), to analyze which processes and parameterizations cause differences between models. To this end, the authors perform a sensitivity analysis, where five different factors (CO2, population density, land use, lightning and climate) are individually modified. The authors identify land use as the most important factor for differences between models and discuss several potential routes to improve global fire models. The manuscript represents a significant contribution to attempts to improve the parameterizations of Earth system models. It is well written and relatively easy to understand. I have, however, one major concern regarding the setup of the sensitivity analysis, which also effects a part of the findings presented in the manuscript (see comments below). This point should be accounted for before submitting a revised version.

General comments:
In my opinion, the design of the sensitivity analysis is not sufficient to support all conclusions made in the manuscript. The setup is suitable to analyze differences between models with respect to one factor (e.g. CO2). This is the case, because the modification of the factor (e.g. keep at constant value) is the same for all models, so differences between models have to result from the shape of the relation between

this factor and the examined variable, burned area, which is implemented in the model. This is nicely explored in the manuscript by additional analyses of how the respective factors affect processes in the model. However, the setup is not suitable to compare the relative effects, meaning the relative importance, of different factors, e.g. population dynamics and climate. The reason is that the factors show trends of different strength over the examined period (1900-2013). It is not clear to me how the authors separate the effect of the trend from the effect of the relation between factor and the simulated burned area (see specific comments below). For example, let us assume that both CO2 and climate have a similar effect on burned area in the models. However, CO2 shows a strong trend in the period 1900-2013, while climate does not. This is enhanced in the setup of the sensitivity analysis by choosing a low value for CO2 for the experiment, but average values of climate variables. Consequently, the slope of the relative difference in burned area (e.g. Fig. 2) will be larger for CO2 than for climate, although both factors are (hypothetically) equally important in the model. This also affects the relative differences between models: If the general effect of CO2 is amplified compared to climate in our hypothetical case, also the differences between models will be larger for CO2 than for climate. The authors need to clarify this, both in the methods and discussion section of the manuscript.

1) *We thank the reviewer for their assessment and the acknowledgement of our contributions. We address the methodological concerns by three points:*
    - *We agree with the reviewer that we do not separate the effect of the trend in the driver from the effect of the relation between factor and simulated burned area. We used the word term sensitivity loosely to mean the net response to the forcing, while the reviewer interprets it more formally as a change in response variable per unit change in forcing. To avoid confusion we adopt the reviewer's definition and thus have changed the title to "Response of simulated burned area to historical changes in environmental and anthropogenic factors: A comparison of seven fire models". As our goal was to understand which factors cause the response of burned area over the historical period we therefore need to look at the response given the present trends. Finding a high sensitivity for a forcing factor that has no trend would not directly help to understand the response over the historical period. We now reword the appropriate text passages accordingly and address which factors influenced the burned area over the historical period. Further, we highlight that response in burned area are caused by both: the sensitivity of the model and the imposed trend in the forcing. We also add the trends of the forcing datasets in the table 4 and include three sentences 'Response of simulated burned area to individual drivers' section:*

The population density forcing dataset has the strongest trend in the relative differences between the transient forcing and the year 1920 value followed by the land-use and land cover change dataset. The trend in atmospheric CO2 concentration is higher than the trend in the lightning dataset, which is more than twice as strong as in the air temperature. Wind speed shows the lowest trend of all investigated driving factors (see tab. 4).

- *The reviewer notes that we use an average of the climate variables. This is not exactly what we did. We recycle the 20 first years that are available as climatic forcing (1900-1920) in the climate sensitivity simulations. However the reviewer is right that due to this there is no difference between the reference and the sensitivity simulation in the first 20 years of our comparison. We therefore now compute the trends of the in burned area between reference and sensitivity simulation starting in 1920 until the end of the simulation (2013). As we investigate the trend of differences with a consistent starting point for all factors (not simply the differences between sensitivity and reference simulation) we can now also compare the importance between the factors for the simulated historical changes of burned area.*

  We add in the manuscript in the Methods:
  The resulting difference in burned area between the simulations is then a combination of the changes in the forcing and the sensitivity of the model to that forcing factor.

  and in the Response of simulated burned area to individual drivers section (see also reply 21):
  The response of burned area to the individual factors is determined by the changes in the driving factors and the sensitivity of the model to these changes.

We use the word sensitivity now only in these places and for "sensitivity experiment". In other places sensitivity has been replaced with "response of simulated burned area to" .

- *As a second change we now use the absolute differences instead of relative differences. As the CO2 concentration for instance was fixed at the value of 1750, for some models the burned area that is used to normalized is much smaller than it would be if the value was set to the value of 1900. All models have a comparable magnitude of burned area for present day therefore the absolute changes are also comparable and the comparison between models is not strongly influenced. The reviewer did not directly request this but we think that*

this increases the comparability between the factors. Our conclusions are not affected by this change but the quantification of trends is more meaningful. We add in the Methods section
Two of the models (CLASS--CTEM and CLM) started the simulations later than the others (1861 and 1850, respectively) and due to limitations in data availability the reference year of the forcings used in the spin-up varies (see tab. 1). We account for these differences in starting years between models and of the forcing factors by limiting our analysis to the period where all factors are different from the ones used in the spin-up (after 1921). These differences still influence the absolute differences, we therefore quantify the strength of the impact through the slope of a regression line and do not interpret the offset.

Specific comments:
P 2 L 7 Please replace 'regularly' by a more detailed description, such as 'at least once in 100 years' or similar. Does that mean that at least 60% of the land surface are never affected by fire?

2) The descriptions in the literature were not hat precise, thus we have removed the sentence.

P 2 L 12 Please put the 5.6 ppm CO2 into context: Which percentage of the total feedback per degree of warming does this correspond to?

3) We now include the strength of the global land climate-carbon-cycle feedback (17.5 ppm K-1) as a context. It corresponds to a percentage of approximately 32%.
Analyses based on observations of the pre-industrial period suggest that the contribution of fire to the overall climate–carbon-cycle feedback is substantial with 5.6 ± 3.2 ppm K-1 CO2 (Harrison et al., 2018) while the strength of the global land climate–carbon-cycle feedback estimated from Earth system simulations (Arora et al., 2013) is 17.5 ppm K-1 (Harrison et al., 2018). However, comparing potential fire-induced losses from terrestrial carbon pools and stocks of solid pyrogenic carbon in soils and ocean, fire may also be a net sink of carbon and Earth system simulations show a negative effect of fire on radiative forcing (Lasslop et al., 2019).

P 2 L 26 Please explain the term 'woody thickening' shortly. How does vegetation composition change?

4) We modified the manuscript as follows:
It can lead to an increase in the abundance of woody plants ('woody thickening'; Wigley et al., 2010; Bond and Midgley, 2012; Buitenwerf et al., 2012) [...]

P 2 L 28 Why does reduced stomata conductance lead to increased fuel moisture? Is it assumed that plants take up water from the litter layer? Please explain this shortly.

5) It is assumed that the water saving increases soil moisture and in consequence fuel moisture, including the living biomass contribution to the fuel load and the amount of litter on the soil surface.
On the other hand, decreased stomatal conductance and lower transpiration can lead to enhanced water conservation in plants. This increases the moisture content of soil as well as vegetation moisture content and consequently live and dead fuel moisture contents, which decreases flammability and in consequence reduces burned area.

P 3 L 6 It is quite difficult to understand this sentence. Please start with the end (nr offires times size) and may be split into two sentences.

6) We rephrased the sentence:
Burned area can be expressed as the number of fires multiplied by their fire size. The increase in burned area due to changes in ignitions is expected to differ between regions with varying population density as the largest fires occur in unpopulated areas (Hantson et al., 2015a).

P 4 L 21 Does the around 150 year shorter spin-up for two of the models have effects on the fuel amount? Or is the turnover of the fuel fast enough to exclude that the models with shorter spin-up have less fuel?

7) The described simulations start from a spinup simulation where carbon pools were equilibrated. We add a sentence to describe this point in the Methods section:
The baseline FireMIP experiment (SF1) is a transient simulation from 1700-2013, in which atmospheric CO2 concentration, population density, land-use, lightning, and climate change through time according to prescribed datasets. The baseline and sensitivity simulations start from the end of a spin-up simulation with equilibrated carbon pools (see Rabin et al. (2017a) for details of the experimental protocol).

P 5 Tab1 Why are only low values of CO2, population density and land use(?) included in the sensitivity analysis? Would it not make more sense to either use intermediate values, similar to climate and lightning, or, alternatively, test high values in addition to the low ones?

8) See also reply 1. The experiments were designed to understand the influence of the historical variation in the driving factors on the simulated burned area. Therefore all factors were individually held constant at the initial conditions, e.g. the conditions that were used in the spin-up. Lightning and

climate varied in the historical baseline simulation from 1900 and were set to the first twenty years before, as no forcing dataset is available before that time and because the interannual variability in climate is important (so using only one year is not an option). We now compute the trends starting with the year 1920, when all factors vary. Results may be slightly different when fixing the forcing at values of different years, but as we are interested in how the historical changes influenced the historical simulations in burned area we think the interpretation of the high values would be less direct. The sensitivity simulations now start with a state that existed in the past (neglecting, of course, any existing errors in the models and forcing datasets). Starting the simulation with the high values would be a hypothetical case, as the models also slightly depend on their history. Technically this would also mean that the sensitivity simulations all require a separate spin-up. They would start from different initial conditions and although they would end with the same forcing the model state would likely be different as for present day ecosystems are not in equilibrium due to global change.

P 6 L 11 Please add a short description of how these data sets differ, beyond the retrieval algorithms, since this is important to understand the results (e.g. agricultural fires in GFED4s)

9) We now include an improved description how these datasets differ.

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.

P 6 L 16 In which direction is the distribution skewed? Does the model resolution have an effect on the shape of the distribution?

10) The distribution of burned area has a very large fraction of 0 and small burned area, high fractions of burned area have a very low frequency. We add a plot indicating the influence of individual datapoint in the comparison between GFED4 and FireCCI50 in the supplement. Without transformation a very small fraction of the data points determines the correlation, 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. As the correlation should provide a global evaluation of the model a much higher influence of individual grid cells is not desirable. As the models are all aggregated to the same spatial resolution the model resolution does not have an influence on the distribution.

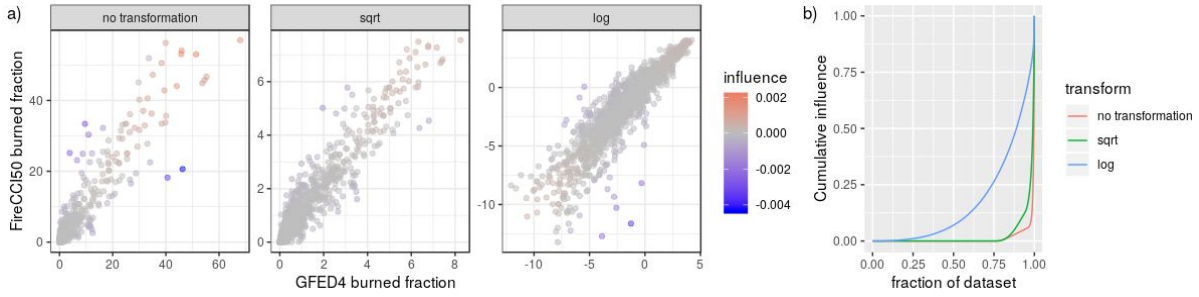

Figure A9: Scatter plots for the GFED4 and FireCCI50 dataset without transformation, square root transformation and log transformation (a), the color

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.

P 6 L 21 The values 0.01 and 0.2 refer to the GFED4 and FireCCI50 data sets, I assume? Please make this clear.

11) We clarify in the manuscript
[...] yields uncertainty estimates of 0.01% (GFED4) and 0.2% (Fire CCI50)

P 8 L 9 - P 9 L 2 I think this part should be shifted to the discussion.

12) We did not separate Results and Discussion but directly discuss the results following the presentation. We shortened the indicated paragraphs slightly to have more emphasis on the results and moved part of it to the "Implications for model development and applications" section.

P 9 L 4ff I do not understand the line of argument: In the first three experiments (CO2,population,land use), relatively strong trends and large model differences throughout the 20th century are reported. In the other two experiments, the trends are weaker. However, this result may be influenced from the setup of the sensitivity analysis, since there are trends in CO2, land use and population density over the 20th century. Population density, for instance, is kept at the low value of 1900 in the experiment, so it is logical that the rel. diff. BA increases over the 20th century for models, which assume a positive effect of population density on BA (e.g. LPJ-GUESS-SPITFIRE), due to the trend in population density. For models which

assume a negative effect of population density on BA (e.g. LPJ-GUESS-SIMFIRE-BLAZE), the opposite is the case.
However, it is not described how the effect of the trends (e.g. increase in population density) is separated from the effect of the factor in the model (e.g. effect of population density on fire).

13) See also reply 1). Population density is kept at the value of 1700. We now use the absolute differences. The initial values of land use, CO2 and climate stem from different years. This is because climate data were only available from 1900 onwards. We now compute the trends starting in 1920 when all factors vary, with low influence on the results. Fig. 2 already showed the strong interannual variability of climate and lightning and the absence of trends over the whole period. Qualitatively the spread between models for population density is logical considering the different assumptions in the models, but note that most models assume a curve with a maximum and therefore include positive and negative effects. Quantification of the net effect and also the magnitude of the effect therefore requires the sensitivity simulations provided in this study. As we aim to quantify the effect of forcing factors over the simulation period we quantify the response in burned area given the historical trend. Quantification of the burned area response with a hypothetical trend (for instance a doubling) would not allow to understand the historical simulated trends.

Figure 2 and Table 4 are only suitable to compare the relative effect of one factor between models, but not the relative importance of different factors. Maybe the relations between rel.diff. BA and lightning, and also rel.diff. BA and climate, are weak because the trends over the 20th century are not as pronounced as for the other factors, and also average values (1901-1920) are used for the experiments. In this case, the mean values of baseline scenario and the experiments would be very similar to each other, and variations would be randomly distributed over the 20th century, which is partly consistent with Fig. 2. Therefore, I am not convinced that the slope of the rel. diff. BA over the 20th century (Tab 4, Fig 2) is a good measure of the strength or importance of a certain factor in the model, compared to other factors.

14) We now use the absolute differences, see reply 1. We assume this may also again relate to the fact that we did not separate out the strength of the trend in the driving factor. See previous comment and reply 1 and 8. We now clarify that we are interested to understand which factors cause the simulated trends over the historical period. Note that the climate was not averaged over the 1900-1920 period but recycled. We now compute the trends for the absolute differences and for the period 1920 to 2013 for which all factors vary.

P 12 L 11 Please add ’concentrations,’ after ’CO2’.

15) We replaced all occurrences of 'CO2' with 'atmospheric CO2 concentration' to be precise.

P 16 L 3 Please explain shortly why the presence of lightning always leads to a net suppression of fire by humans.

16) The effect of increasing human ignitions is strongest if no other ignitions are present. If lightning already ignited a fire and additional human ignition has little effect. This was tested with the CTEM model, which is also part of this intercomparison study. We include in the text:
The presence of lightning ignitions reduces the limiting effect of a lack of human ignitions on burned area. For the CLASS-CTEM model as soon as lightning ignitions are present, the net effect of humans is to suppress fires, even though the underlying relationship assumes an increase in ignitions with population density (Arora and Melton, 2018, supplement). This may explain why global models assuming an increase of ignitions with increases in population density are able to capture the burned area variation along population density gradients (Lasslop and Kloster, 2017; Arora and Melton, 2018) and why global statistical analysis find a net human suppression also for low population density (Bistinas et al., 2014).

P 18 L 15ff From the listed parameters, only the first two (precipitation and temperature) are climate variables. The others are dependent variables, which are also influenced by other factors (e.g. CO2). Please explain why you include them in the test. Moreover,I would like to see an analysis of the effects of wind speed. Is there a trend in wind speed from 1900 to 2013 ?

17) *We include the vegetation parameters in addition to the climate parameters as climate influences fire not only directly but also through its influence on vegetation. We modify the included explanation: "The influence of climate on burned area is complex; it influences burned area through the meteorological conditions and through effects on vegetation conditions that influence fuel load and fuel characteristics (Scott et al., 2014). We therefore correlated for each grid cell changes in physical parameters (precipitation, temperature, wind speed and soil moisture) and vegetation parameters (litter, vegetation carbon and grass biomass) with changes in burned area."*
Note that CO2 is not different between the simulations compared here, only climate differs. In addition, we add the linear regression slope and the standard deviation for wind speed in table 4; over 1921 - 2013, the relative difference in wind speed has a significant negative linear regression slope (-0.012 +- 0.006). We add 'Wind speed shows the lowest trend of all investigated driving factors (see tab. 4).'

P 18 L 30 The word 'is' occurs one time too often.

18) Removed.

P 19 L 10-12 I am not sure that this statement is valid, given my concerns on the setup of the sensitivity analysis above.

19) See reply 1, 8, 13, 14. This refers to "Representing human influence on fire is the major challenge for long-term projections. 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."

We have now improved the computation of trends. To assess the importance of certain factors in trajectories the underlying trend is important, a separation of the trend in forcing from the sensitivity of the model would therefore not improve the assessment. However changes in the trends of the forcing factors for future can change the results we therefore included:

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.

P 19 L 32 The word 'Table' is missing in the brackets.

20) It is included now.

P 21 L 14 How strong is the trend in changing climate compared to other trends, e.g.population density and CO2?

21) We now quantify the trends in the forcing factors. It is however questionable how comparable these changes are between factors. Also the global increases in CO2 are more meaningful than global changes in temperature as CO2 is fairly similar in different locations while the changes in temperature vary regionally. For text modifications, see reply 1.