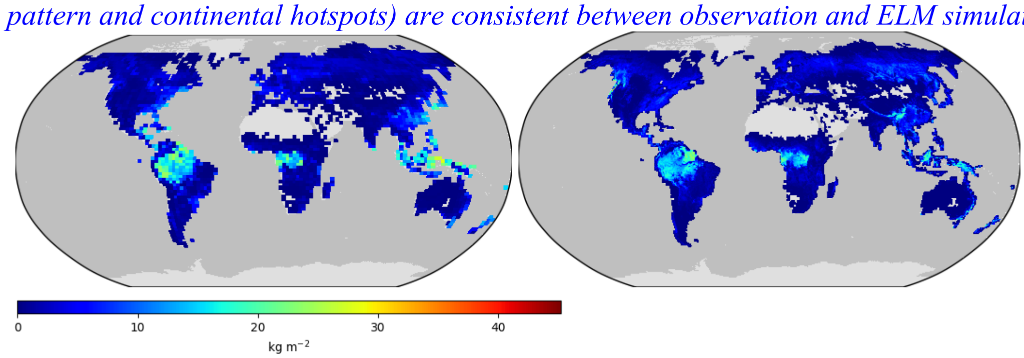
Dear Editor and reviewer,

We very much appreciate the reviewers' comments and feel that they have allowed us to substantially improve our manuscript. Below, we repeat the reviewers' comments and then respond to each comment individually in *blue italics*. Related modifications in the revised manuscript are highlighted in red.

Reviewer #1 (accepted as is)
Response:
We appreciate the reviewer's efforts on reviewing and improving this manuscript.

Reviewer #2
1. I previously asked how well the simulated biomass captures observed biomass(e.g. remote sensing estmiates) as any biases in modelled biomass will cause errors in the simulation of fire. The authors responded to this question by comparing a global total estimate of biomass (Figure S6). This comparison is not meaningful because a global total biomass can originate from various regional patterns of biomass. I request to make a proper evaluation of the simulated biomass by e.g. comparing a map of simulated biomass with a map of observed biomass and making a difference map (e.g. Biomass CCI dataset). This will allow assessing if regional biases in simulated biomass (and hence other fuel properties) might cause biases in the simulation of burned area. Ultimately, it is completley unclear how such an differences would affect the simulations in a coupled model.

Response:
We compared spatial distribution of aboveground living biomass form GEOCARBON dataset (left) and ELM simulation (right) for present-day condition. We found that the spatial distribution (latitudinal pattern and continental hotspots) are consistent between observation and ELM simulation.



2. In the response it is written that "we tuned the DNN-Fire surrogate model towards ensemble mean with standard deviation across 14 GFED regions" which implies that the standard deviation was considered in the tuning. However, the standard deviation of burned area is not included in the used cost function in equation 8. This needs to be clarified.
Response:
We have clarified in the main text that the DNN-Fire surrogate model was tuned towards ensemble mean across 14 GFED regions (line 173).

3. Figures 4 and S7: Some of the symbols/colours are used for two regions and hence cannot be distinguished (e.g. BONA and CEAS). The colours, symbols and legends need to be revised.
Response:
Figure 4 and S7 are updated to use different colors for different regions.

4. The analysis of the sensitivity of the results to different forcing data is interesting as it reflects actually a strong influence of the forcing data on the simulation result. In order to identify the most reliable

forcing data a comparison with the observed burned area and biomass would be insightful. Which forcing dataset results in model simulation closest to the observations?
Response:
In terms of climate forcings, GSWP3 dataset has been demonstrated to be the best climate dataset and yield the best model performance compared with observations for carbon, water and energy cycles (Zhu et al., 2019).

Figure S5: To make the plot more clear, it would be usefull to include the 1:1 line each panel.
Response:
1:1 line is added in the Figure S5.

Lines 360-370 are not clearly written and some sentences are repetitive. Please revise.
Response:
We have updated line 360-370, and remove duplicated sentences.

5. Figure 7 looks fuzzy and the colours are difficult to see.
Response:
Figure 7 is updated in terms of line width and colors.

6. Lines 101-102: "Although explicit processes are simulated, the accuracy of process-based wildfire models are highly dependent on parameterization, which is computationally expensive" - This statement is mis-leading as indeed most process-based wildfire models like the models within FireMIP actually were never parametrised using computational approaches. Parameters were rather taken from literature sources during model development. A proper calibration of process-oriented models, i.e. by using a cost function and optimization algorithm was rarely done and otherwise the computation time is comparable to the time needed for the building and training of neural network models.
Response:
We tried multiple tuning exercises for ELM process-based wildfire simulations. We found that the parameterization could largely improve the simulation accuracy in terms of burned area. However, the whole tuning exercise required global sensitivity analysis and large model ensemble simulations, which are computationally expensive and time consuming. In the revised manuscript, we have cited relevant paper in order to support the statement of expensive process-based model parameterization (line 103).

7. Finally, I'm not really convinced by the study. Although you can nicely demonstrate that the DNN can 1) reproduce the simulated burned area of the Earth system model and 2) it also captures regional total burned area from observations, the study does not contribute to an improved "understanding of human, climate, and ecosystem controls on fire number, fire size, and burned area" (as motivated in the abstract). As the DNN models are only trained against burned area, no statements about number and size of fires can be done. Furthermore, the ecosystem controls are mostly represented by the input variables tree cover and biomass. As tree cover is always described from an input dataset and biomass is used from the base simulation, the DNN models are actually inconsistent because they have been trained against different sources and magntiude of burned area (but always using the same magnitude of tree cover and biomass!). As fire has a clear reducing effect on tree cover and biomass (Lasslop et al. 2020), it is obvious that the relation between fire, biomass and tree cover sis actually inconsitent in some DNN models because the spatial patterns of biomass and tree cover do not correspond to the patterns of fire. Hence the approach cannot give meaningful insights on how ecosystem controls affect fire and I am not convinced that this can be transfered to a coupled model. The abstract and discussion needs to be revised in order to make this limiation clear.
Response:
We agree that the offline tuning has uncertainty stemming from tree cover dataset and ELM simulated vegetation biomass. A fully couple DNN-Fire and E3SM land models will overcome such shortcomings

and yield consistent tree cover, biomass, burned area simulation, which is our long-term objective for fire modeling. We will achieve this long-term goal with multiple steps. This study is the first step to develop and tune the standalone wildfire model within the E3SM land model interface so that burned area dynamics could be reasonably simulated. The current study is an important step towards a fully coupled E3SM + DNN-Fire model.

Reviewer #3

1. Zhu et al. present a deep neural network that 1) emulates the fire model used in CLM and 2) optimises on observed burnt area from multiple products. They include an evaluation and the new "surrogate" model and a discussion on potential uses. While there are no flashy results, the DNN approach presented shows great promise in many applications of global fire modelling, and its documentation is a perfect fit for GMD. The m/s is very well written and easy to understand - quite an achievement for such a novel and potentially complicated approach. It is perhaps the most enjoyable read I have seen in GMD.

Response:

We appreciate the reviewer's positive comment.

2. Normality assumption in equation 8. The distribution of burnt area is extremely non-normal, and so should the probability in the model-to-observational error in the cost function. Having said that, the evaluation of the DNN shows it still performs well, so I won't make a new analysis a requirement for the m/s to be accepted. The authors should discuss the limitation of this normality assumption on model use, and possible ways of developing it. (Kelley et al., 2021) for example, uses a zero-inflated logit transformed least squared approach (though that is for a maximum likelihood approach), which might be useful? If your method struggles with non-unimodal cost functions, maybe a simple logit transformation would be better?

Response:

We have discussed the limitation of the normality assumption in the revised manuscript and cited relevant papers that could help resolve this non-normality issue (line 390-393).

3. Lack of biomass feedback in DNN model in evaluation. Presumably, the "base" model still represents vegetation and biomass feedbacks that the new scheme does not? (Correct me and the m/s if I misunderstood). This means 1) evaluation of base model will be at a disadvantage, as vegetation/fuel feedbacks represent another source of non-linear interactions that could exasperate errors in the base model. And 2) it limits the applications for the surrogate model for future climate and environmental change. Excluding biogeochemical-fire feedback is fine for this study (as the authors point out in the discussion, this study is just a first step). Still, there should be some acknowledgement in the model evaluation and discussion.

Response:

Since our DNN-Fire model is trained at an offline mode, the feedback between DNN-fire simulated burned area and biomass is not considered in this study. We have added some sentences to acknowledge such limitations (line 363-365).

4. Line 41/42: (Lasslop et al., 2020) suggests significantly small changes in tree cover across the models in their study (I think 3-25%)

Response:

Reference corrected.

Line 83-85: "Two types of wildfire models are widely used: process-based models and data-driven statistical models." (Hantson et al., 2016) first suggested this distinction, so it should be cited.

Response:

(Hantson et al., 2016) is cited in the revised manuscript. (line 85)

5. Line 127-129: It sounds like you're not making any developments to CLM4.5 fire processes here? If that's the case, can you state that (sorry if you do somewhere and I missed it)
Response:
We have clarified that "without modification on process representation" (line 128-129)

6. Line 134/135: As mentioned in the intro, human activity also fragments the landscape (Kelley et al., 2019; Andela et al., 2017). Is this effect represented in the model?
Response:
Human factors are represented in the DNN model. Table 1 summarize that population density is considered for ignition process, also Gross Domestic Productivity (GDP) and population density are related to fire suppression activities.

7. Line 140-141: "shaped region controlled by wind speed and fuel wetness…" is this a rate of spread model such as (Rothermel, 1972)? If so, please cite
Response:
Rothermel, 1972 is cited in the revised manuscript (line 141).

8. Line 173-176: on taking the mean of the observations. I wonder if there's a way of capturing the uncertainty from the disagreement between products using your approach? (I'm not suggesting new analysis, but maybe a discussion point for future work)
Response:
In the revised manuscript, we have discussed the uncertainty in burned area products (line 181-184)

9. Line 198-199: Would it work if you swapped around and used 20% for training and 80% for testing?
Response:
80%:20% (training:testing) sample split rule has been widely used, tested, and demonstrated to be effective for many other machine learning research. We therefore adopt this 80%:20% split rule. The goal is to maximally learn patterns and relationships from the data, therefore, only using 20% of the data for training may lead to a non-satisfactory DNN-Fire model.

10. Equation 8: is the "i" each gridcell?
Response:
We have add "i represents different gridcell" in the revised manuscript (line 224).

11. Line 230-232: How do these numbers compare to fireMIP model results in (Lasslop et al., 2020; Hantson et al., 2020)
Response:
FireMIP models simulated burned area have much large ranges of global burned area from 39 Mha yr^{-1} (LPJ–GUESS–GlobFIRM) to 536 Mha yr^{-1} (CLASS–CTEM), compared BASE-Fire.

12. Line 358/359: "This study is an important step towards fully coupling E3SM and the DNN-Fire models in the future." I'd be interested to see how this progresses, so feel free to suggest me as a review in the future.
Response:
We appreciate the reviewer's positive comment.

13. Line 362-365: Very good point!
Response:

We appreciate the reviewer's positive comment.

14. Line 382-383: "will produce a model dominated by gridcells that have high burned area (e.g., Africa)." using a logit transformation on your cost function might help with this.

Response:

We really appreciate this comment. We discussed the limitation of normality assumption and its potential impact on model performance in line 394-397.

15. Line 385-391: I quite like this pragmatic regional approach, but as the m/s points out, GFED regions were designated partly on present-day bioclimate and fire conditions. Future shifts in bioclimate (i.e. vegetation transitions, climate extremes etc., which are more likely at extreme future scenarios (Swaminathan et al., 2021; Burton et al., 2021)), may impact the performance of future projections given this regional optimisation. A brief comment here on this impact would be great.

Response:

In the revised manuscript, we have add to highlight the limitation of using GFED regions for future simulations (Line 401-403).