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
Zhu et al. develop a machine learning (ML) burnt area model that can be used in place of a process-based algorithm in ELM. This approach was first used to surrogate the fire model of Li et al. which was in CLM (and then now ELM). The ML approach uses a deep neural network to reproduce the process model result (they call it Base). Then by altering the parameters they tuned it to match GFED4 burned area. The paper is clearly written and results are generally well presented. I found the work interesting as this is an important problem. Present process-based fire models are not overly skillful. Much of this stems from the many complexities of fire modelling - especially anthropogenic influences. I am optimistic this paper can be published but I would like to see some careful consideration of my comments below. At present the manuscript is what I would consider an absolute bare minimum of what can be published and there are many opportunities to make this paper into a much better resource to the community. This particular approach could be valuable but I think it needs some expansion to demonstrate how useful imbedding ML approaches in process models can be. As a result I would like to see some expansion of the work to better demonstrate the utility of the approach.

Response:
We appreciate the reviewer's positive comments. We have addressed all major and specific comments below.

1 The DNN-Fire model was subsequently tuned to match GFEDv4 but this is not the only burned area product available (e.g. Chuvieco et al. 2019). Indeed there are many other products now available and they don't agree so well (e.g. Padilla et al. 2015, Humber et al. 2019). I worry that by tuning the model to reproduce one dataset you may get a result closer to that dataset but at the expense of adopting its same biases and thereby potentially not getting as admirable advances in accuracy at it seems. Why not consider all of the available burned area products to produce a burned area estimate that could then be less biased by a single dataset? As, in reality, we are most interested in increasing our predictive skill - not just reproducing an observation.

Response:
We agree that considering multiple existing datasets of burned area could avoid over-parameterization to any individual dataset and thus reduce the DNN-Fire model prediction uncertainty. In the revised version, we considered five prevailing burned area products including the GFEDv4s, FIRE_CCI51, FIRE_CCIT11, MCD64, Fire_Atlas. Comparing the five prevailing burned area products (Table S1), long term averaged

burned area ranged from 424 Mha yr^{-1} to 484 Mha yr^{-1}, and most of the data discrepancy was located over tropical regions (Figure 2). Compared with multi-datasets mean, the Base-Fire model (ELMv1 process-based wildfire model) still had large biases across tropics, temperate, and boreal regions (Figure 2).

In order to make use of the five datasets and reduce DNN model uncertainty associated with over-parameterization towards any individual datasets, we first calculated ensemble mean and standard deviation of the five burned area datasets for each gridcells, then we tuned the DNN-Fire surrogate model towards ensemble mean with standard deviation across 14 GFED regions. All new results were updated throughout the paper (highlighted in the manuscript with red color).

***Table S1.** Burned area datasets used in this study*

Dataset name	*Temporal range*	*Spatial resolution*	*Global burned area, mean (std)*	*Citations*
GFEDv4s	*1997-2015*	*0.25 degree*	*455(39)*	*(van Der Werf, Randerson et al. 2017)*
Fire_CCI51	*2001-2019*	*0.25 degree*	*476(26)*	*(Lizundia-Loiola, Otón et al. 2020)*
Fire_CCILT 11	*1982-2018*	*0.25 degree*	*484(20)*	*(Lizundia-Loiola, Pettinari et al. 2018)*
MCD64	*2001-2019*	*0.25 degree*	*424(35)*	*(Giglio, Boschetti et al. 2018)*
Fire_Atlas	*2003-2016*	*0.25 degree*	*459(43)*	*(Andela, Morton et al. 2019)*

Note: the long-term average global burned area was calculated using data with the same overlapping temporal range (2003-2015), unit Mha yr^{-1}

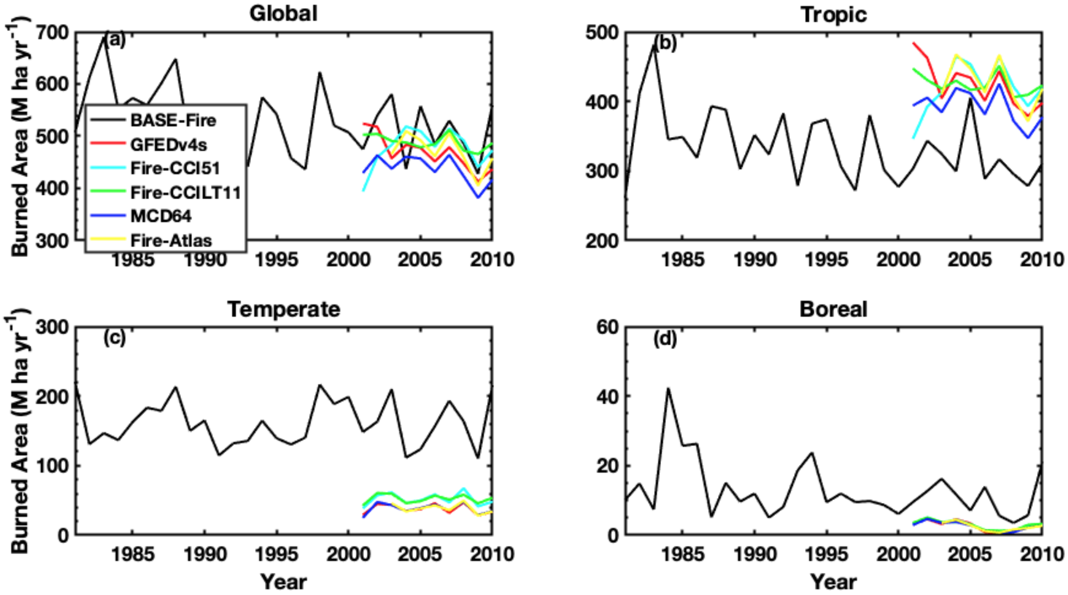

Figure 2. *BASE-Fire simulated and burned area datasets of GFEDv4s, Fire-CCI51, Fire-CCILT11, MCD64, Fire-Atlas. (a) Global scale; (b) Tropical (S23.5° -N23.5°); (c) Temperate (N23.5° - N 67.5°); and (d) Boreal (north of N 67.5°) regions.*

2 By surrogating Base-Fire, the DNN-Fire then integrates/assumes the biases and issues apparent in ELM's simulations (e.g. too much/little biomass, too dry/wet soil, etc.) and produces a model that aims to get the right result (burned area matching GFED) potentially for the wrong reasons (based on biased inputs). Why not run an ensemble approach with different forcing datasets (e.g. met forcing of CRUJRA in addition to GSWP3, or a different land cover (if using prescribed), etc.) to try and give at least a measure of the uncertainty in these inputs to the DNN? We have found for our model (run in normal process-based mode) the results can be surprising and have some strong impacts for certain variables. Gitta Lasslop looked at this too and found a large impact upon fire, primarily due to the wind speed differences (e.g. Fig 3 in Lasslop et al. 2014). Alternatively using an observation-based product of one of the ELM variables (Table 1) like soil wetness or above ground biomass as another means to look at the influence of input bias.

Response:

We agree that the model uncertainties from biased inputs are potentially important. Therefore, we investigated the DNN-Fire model uncertainties from 1) surface climate; 2) soil moisture inputs; 3) interactions between climate and soil moisture. Unfortunately, the uncertainty from biomass (fuel load) was not evaluated, due to lack of 2001-2010 transient data for global vegetation biomass.

For climate forcing uncertainty we drived the DNN model with (CRU-JRA (Onogi 2007), NCEP-DOE2 (Kanamitsu 2002), in addition to the default GSWP3 (Dirmeyer et al., 2006). For other ELM input variables, we evaluated soil moisture uncertainty by driving *the DNN model with the NOAA NCEP-NCAR-CDAS-1 (Kalnay 1996) topsoil moisture product.*

Overall, climate forcing was a big uncertainty source for burned area simulations. For example, over the three largest fire regions (SHSA, NHAF, SHAF), major uncertainty came from climate forcing rather than topsoil moisture (Figure S3). Furthermore, among the three climate forcings, CRU-JRA was close to the default GSWP3 forcings, while NCEP-DOE2 forcing led to large reduction in simulated burned area.

In the revised manuscript, we add a paragraph to discuss the potential uncertainties from input variables (Line 360-375):
"We acknowledge several challenges and limitations in our modeling framework. First, the DNN model uncertainty was subject to the accuracy of climate forcings as well as other physical driving variables simulated by the physical wildfire model (ELMv1). For example, in addition to the default GSWP3 climate forcings dataset used in the study, CRU-JRA [Onogi et al., 2007] and NCEP-DOE2 [Kanamitsu et al., 2002] reanalysis forcings were also widely used and potentially different from GSWP3 forcings. ELMv1 used climate forcing (e.g., temperature, precipitation, wind speed, relative humidity) to simulate soil temperature, soil moisture, fuel load and so on. These simulated variables served as inputs for the DNN model and could also result in prediction uncertainty. It was challenging to eliminate the forcing uncertainties in this work, but we could at least evaluate the magnitude of these uncertainties. We ran the DNN-Fire-OBS model with alternative forcings of CRU-JRA, NCEP-DOE2, and CDAS soil moisture from 2001 to 2010 and compared the results with DNN-Fire-OBS driven by default inputs (GSWP3 climate and ELMv1 simulated soil moisture) (Figure S3). The results showed relatively larger uncertainties from climate forcing than that from soil moisture forcing particularly over the major fire regions (e.g., SHSA, SHAF, and NHAF). Future work will focus on evaluating the uncertainties from fuel load and fuel temperature variables."

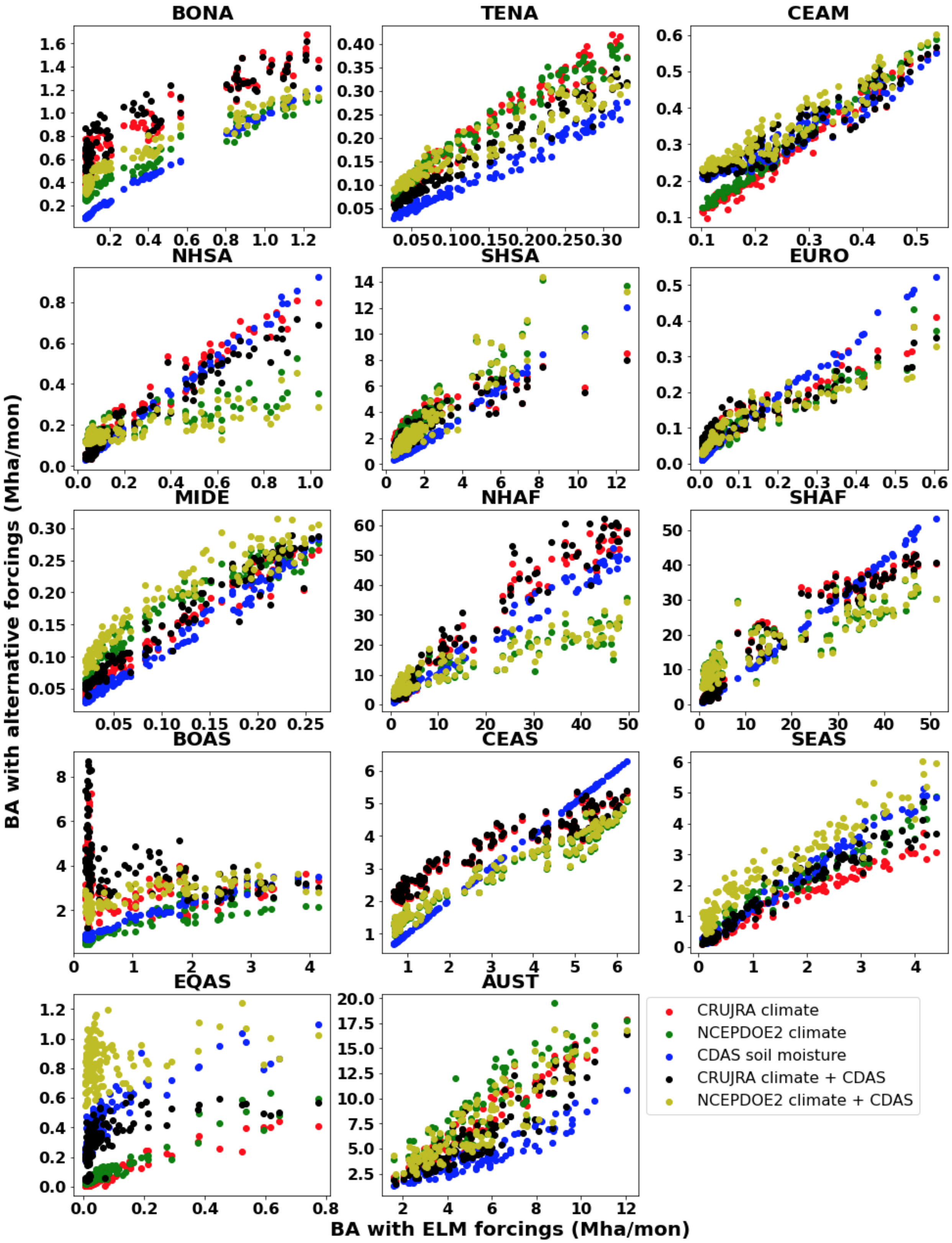

Figure S3. *Sensitivity of modeled burned area (2001-2010 long-term averaged) to climate forcings (including temperature, precipitation, wind speed, relative humidity) and*

soil moisture. X-axis was burned area simulated by the default model using GSWP3 climate forcing and ELMv1 simulated soil moisture. Y-axis were models with alternative climate forcing (CRUJRA, NCEPDOE2) and soil moisture product (NCEP CDAS soil moisture).

3 Around line 188 you describe the training/testing split. This approach of doing it randomly makes me wonder if the influence of spatial autocorrealtion will result in an overly optimistic error estimate. Especially as fire is likely autocorrelated. There are many papers in the literature discussing the dangers of random sampling on spatially correlated data (e.g. Roberts et al. 2017; Meyer et al. 2019; Ploton et al. 2020; Kühn and Dormann, 2012). I would suggest an alternate strategy be employed. It also wasn't clear how this test/train split results were integrated. I think it was just in the model score?

Response:

In the revised manuscript, we used "stratified random sample method" to maximally eliminate the impacts of spatial autocorrelation on random sampling [Wang 2012]. The burned area over all grid cells were first divided into three subgroups or "strata" based on the magnitude of the burn (low burn 0-33 percentile, medium burn 34-66 percentile, high burn 67-100 percentile). Then the grid cells were randomly sampled, but with the constraint that samples were drawn from each strata according to the ratios of samples within each strata. In this case, the spatially correlated gridcells (e.g., nearby highly burned gridcells) were more likely divided into different datasets of training/testing, compared with the straightforward random sample method.

In the revised manuscript, we add a paragraph to describe and discuss the stratified random sampling approach (Line 193-199):
"Furthermore, the random sampling was stratified in order to reduce the risk of sampling, e.g., adjacent high fire grid cells. All grid cells were first divided into three "strata": low burn (0-33% percentile), median burn (33%-66% percentile), and high burn (67-100% percentile) grid cells based on the magnitude of the burn. The stratified random sample assured the sampled grid cells for training and testing had the same ratios of low/medium/high burn, thus eliminating the sampling bias from spatial autocorrelation [Wang et al., 2012]."

4 What is the impact of training on such a short timeseries of fire observations when some regions have fire return intervals of >100 years? Also how representative are those years chosen? Would it matter if you instead trained on 2006 - 2015 and tested on 2001 - 2005?

Response:

We agree that the fire return interval could be longer than the observation period (2001-2010 in this study). And the fire return interval may impact the modeling of site level fire *dynamics. We argue that across a large scale such impact will decrease due to spatial heterogeneity of the fire occurrence (gridcells have the same fire return interval, but with fire occurring in different years).*

In order to assess the representativeness of the year chosen for training and testing, we trained and evaluated model performance with selected year of test datasets 1) 2001-2002, 2) 2003-2004, 3) 2005-2006, 4) 2007-2008, 5) 2009-2010. The rests were used as training datasets. It resulted in five different models, each trained by 8 years of data; and tested with the remaining 2 years of data. We found that the selection of training or testing years did not significantly change the model performance (Figure S1).

In the revised manuscript, we add a section to describe and discuss the impacts of selected year of test datasets on model performance (Line 199-205):
"In addition to random sampling, we also investigated the impacts of data choice on the model performance, by sampling the testing datasets within specific years (e.g., 2001-2002, 2003-2004, 2005-2006, 2007-2008, 2009-2010) and used the rest of the years for training. We found neglected differences among the models (Figure S1) indicating the choice of training/testing data years were not impactful. Therefore, we will discuss the results with stratified random sampling approach as the major results throughout the paper."

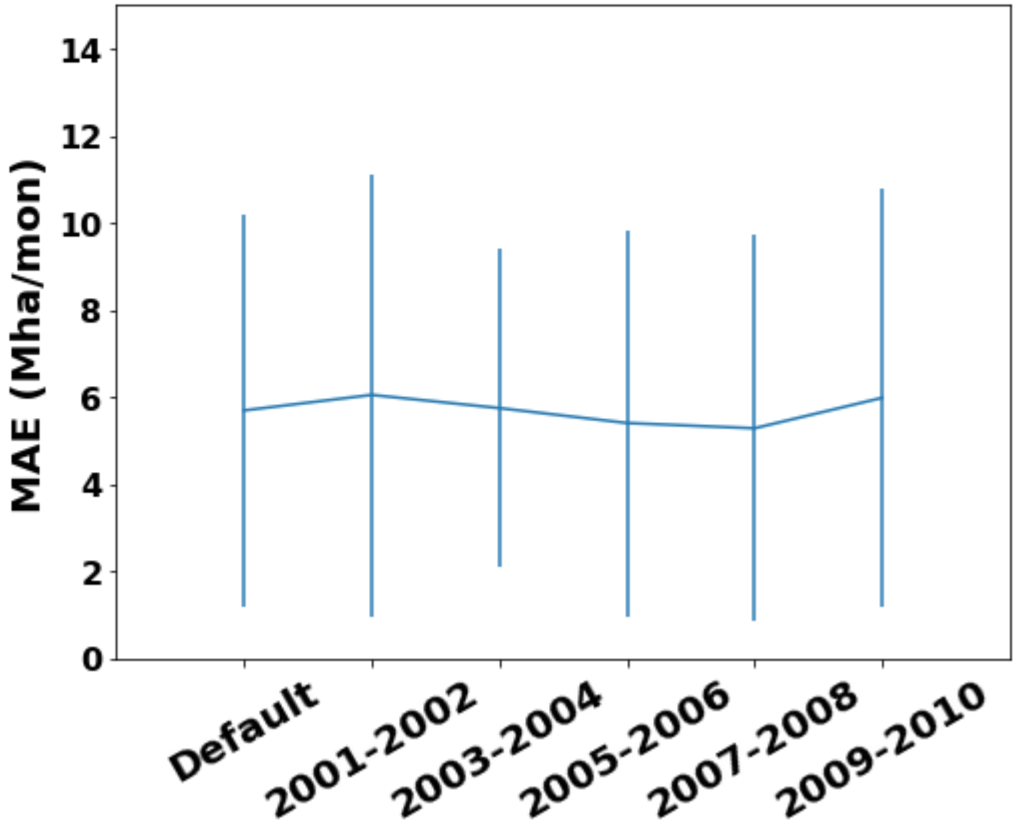

***Figure S1**. Model performance evaluated with testing datasets of default (20% randomly selected samples), or fixed to 2001-2002 period, 2003-2004 period, 2005-2006 period, 2007-2008 period, and 2009-2010 periods (the rest of the dataset was used as a training dataset.).*

5 Figure 7 is the same as the years trained upon so there is little interesting information here. Basically this is showing that it can do an ok job when tested over the same training region. Why not expand this out beyond the satellite era? How does this do from say 1900 on? Yes there is no satellite data but there are other means to check results (see e.g. Arora and Melton 2018)

Response:

We really appreciate the idea of evaluating model performance during historical periods. In the revised manuscript, we compared the DNN-Fire-OBS model simulated global burned area during 1981-2000 periods against the charcoal index inferred burned area (Arora and Melton 2018). We found that the DNN-Fire-OBS model was able to capture the decadal declining trend of burned area at global scale.

In the revised manuscript, we add several sentences that compared DNN-Fire-OBS with charcoal index inferred burned area (Line 342-346):

“Validation was also conducted for the historical period 1981-2000, when most of the satellite based burned area data were not available. Compared with charcoal index inferred burned area during 1981-2000 (Figure S2), DNN-Fire-OBS model reasonably captured the declining of burned area from ~530 Mha yr^{-1} to 490 Mha yr^{-1} .”

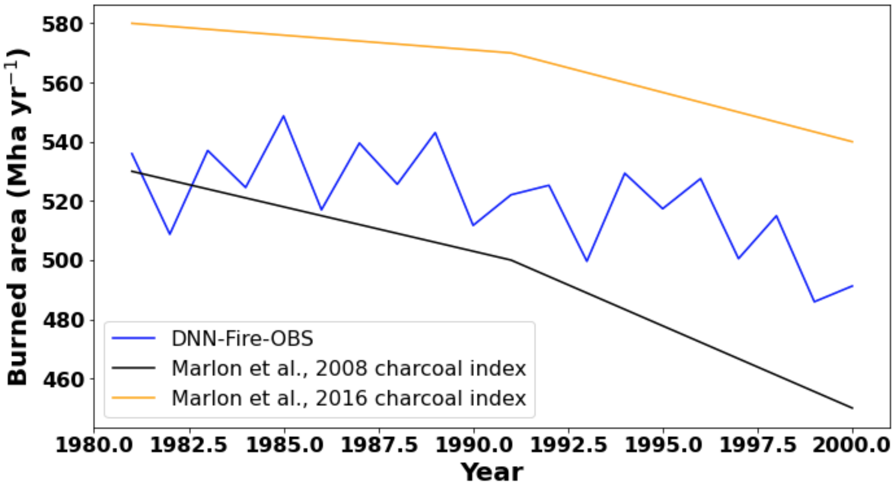

Figure S2. *Comparison of DNN-Fire-OBS model simulated global burned area during 1981-1999 with two charcoal index inferred burned area.*

6 Didn't GFEDv4 offer some uncertainty bounds?

Response:

In the revised version, we used the five datasets average as target variable, and min/max range as uncertainty bounds during training/evaluation. While, the uncertainty of each individual dataset was not accounted during training and testing.

7 Fig 8 to make a stronger demonstration that this is a signifcant improvement, what about plotting the models of FireMIP as further reference points? E.g. Hantson et al. 2020.

Response:

We appreciate the suggestions on comparing DNN model with FireMIP predictions (9 models). FireMIP models simulated burned areas till 2013. Our prognostic simulation period was 2011-2015. Therefore, we took the overlapped 2011-2013 FireMIP results, and compared with observations. We found that FireMIP simulated diverse latitudinal distributions of burned area and generally underperformed compared with DNN-Fire-OBS model (Figure 9), when benchmarked against the averaged latitudinal distribution of the five burned area products.

In the revised manuscript, we add several sentences that discuss FireMIP (Line 339-342):

“We also compared the nine FireMIP models [Rabin et al., 2017; Teckentrup et al., 2018] and found diverse latitudinal distribution of burned area. The across model differences were much larger than the inter-annual variation simulated by each individual model, which indicated large model structural uncertainties.”

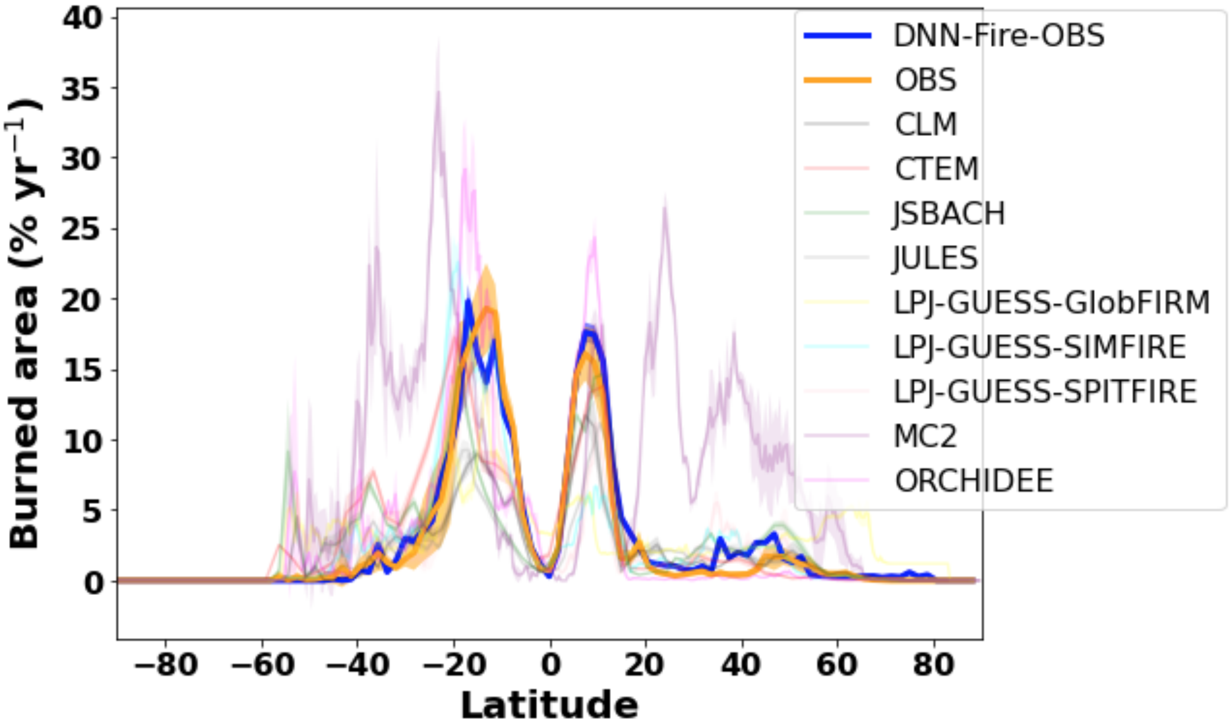

Figure 9. *Prognostic simulation of annual wildfire burned area (2011-2015) with the Deep Neural Network wildfire model fine-tuned with observations (DNN-Fire-OBS) compared with observations and nine FireMIP models outputs.*

8 L41, a more up to date reference would be Lasslop et al. 2020 as it was done with more advanced models

Response:

Citation updated (Line 41)

9 L90: A good reference could be Rabin et al. 2017 as there are some figures showing explictly how the models differ.

Response:

Citation updated (Line 90)

10L186 **-** to be clear, the 14 submodels were combined to produce the global estimates right? Would there be benefit from doing even more sub-regions? What about 20, 50, etc? Where are the diminishing returns here?

Response:

The choice of 14 fire regions was based on the historical convention from Global Fire Emissions Database (GFED) studies. The 14 GFED regions were high level clusters for similar fire behavior, background climatology, and vegetation types. The GFED regions also consider the suitability for comparison with other wildfire studies e.g., atmospheric tracer inversion studies (van der Werf 2006).

We appreciate the reviewer's suggestion of dividing the 14 regions into more sub-regions, which might benefit the model performance. But, we would like to still keep the 14 submodels in this study, for the sake of consistency and easy comparison with other wildfire modeling work.

11 L276 **-** was this talking about the speed of creating DNN-Fire or DNN-Fire-GFED? Several minutes on a laptop? HPC?

Response:

We have clarified the description with "we found that parameterization time could be substantially reduced (several minutes for the global calculation with Intel Xeon Phi Processor 7250 processor)" (Line 297)

12 Fig 8 **-** it seems that DNN-Fire-GFED might be less variable than GFEDv4, is that correct? Is this due to the inputs to the ML or is it a result of the ML approach itself?

Response:

In the revised Figure 8 and Figure 9, the variability was determined by both the interannual variability of each dataset during 2011-2015, also affected by the differences among the five burned area datasets. Therefore, it was expected that the variability of OBS (observations) was larger than the DNN-Fire-OBS model simulated variability, which only accounted for interannual variability.

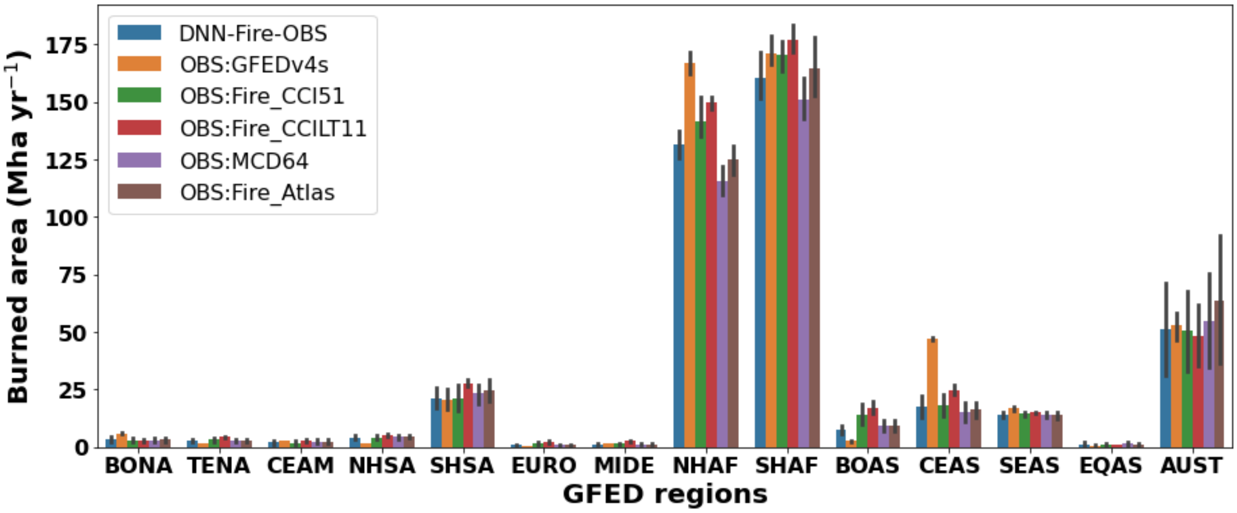

***Figure 8**. Prognostic simulation of annual wildfire burned area (2011-2015) with the Deep Neural Network wildfire model fine-tuned with observations (DNN-Fire-OBS) compared with five observational datasets.*

Reference

Andela, N., D. C. Morton, L. Giglio, R. Paugam, Y. Chen, S. Hantson, G. R. Van Der Werf and J. T. Randerson (2019). "The Global Fire Atlas of individual fire size, duration, speed and direction." Earth System Science Data 11(2): 529-552.

Giglio, L., L. Boschetti, D. P. Roy, M. L. Humber and C. O. Justice (2018). "The Collection 6 MODIS burned area mapping algorithm and product." Remote sensing of environment 217: 72-85.

Lizundia-Loiola, J., G. Otón, R. Ramo and E. Chuvieco (2020). "A spatio-temporal active-fire clustering approach for global burned area mapping at 250 m from MODIS data." Remote Sensing of Environment 236: 111493.

Lizundia-Loiola, J., M. Pettinari, E. Chuvieco, T. Storm and J. Gómez-Dans (2018). ESA CCI ECV Fire Disturbance: Algorithm Theoretical Basis Document-MODIS, version 2.0, Fire_cci_D2.

Van Der Werf, G. R., J. T. Randerson, L. Giglio, T. T. Van Leeuwen, Y. Chen, B. M. Rogers, M. Mu, M. J. Van Marle, D. C. Morton and G. J. Collatz (2017). "Global fire emissions estimates during 1997–2016." Earth System Science Data 9(2): 697-720.

Onogi, K., Tsutsui, J., Koide, H., Sakamoto, M., Kobayashi, S., Hatsushika, H., Matsumoto, T., Yamazaki, N., Kamahori, H., Takahashi, K. and Kadokura, S., 2007. The JRA-25 reanalysis. Journal of the Meteorological Society of Japan. Ser. II, 85(3), pp.369-432.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.K., Hnilo, J.J., Fiorino, M. and Potter, G.L., 2002. Ncep–doe amip-ii reanalysis (r-2). Bulletin of the American Meteorological Society, 83(11), pp.1631-1644.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph. The

NCEP/NCAR 40-Year Reanalysis Project. Bulletin of the American Meteorological Society, March, 1996

Wang, J.F., Stein, A., Gao, B.B. and Ge, Y., 2012. A review of spatial sampling. Spatial Statistics, 2, pp.1-14.

van der Werf, G.R., Randerson, J.T., Giglio, L., Collatz, G.J., Kasibhatla, P.S. and Arellano Jr, A.F., 2006. Interannual variability in global biomass burning emissions from 1997 to 2004. Atmospheric Chemistry and Physics, 6(11), pp.3423-3441.

Arora, V.K. and Melton, J.R., 2018. Reduction in global area burned and wildfire emissions since 1930s enhances carbon uptake by land. Nature communications, 9(1), pp.1-10.

Reviewer #2

This study presents approach to build a deep learning-based model to better simulate burned area as part of an Earth system model. Although several machine learning and data-driven fire models were developed in the last years, this is a first study that directly aims to implement a deep neural network (DNN)-based fire model with a Earth system model. The paper is well written.

Response:

We appreciate the reviewer's positive comments. We have addressed all major and specific comments below.

1 Integration of DNN-based fire model with the Earth system model

The paper is not clear about how the DNN-based model with implemented with the Earth system model (ESM). For the title and abstract, I expect the DNN model was implemented in the ESM. This would allow analyses about how the improved simulation of fire affects the simulated carbon fluxes and stocks in the ESM. But as the paper does not represent such results, I assume that DNN-based fire model was just applied outside of the ESM and that both models were actually not coupled. Hence, I'm wondering how the authors to imagine to couple both models. Especially the final DNN-Fire-GFED setup simulates clearly different burned area then the original BASE-Fire or DNN-Fire models setups. This implies, that for example a much higher simulated burned area in Africa should result also in a much lower biomass in Africa and hence changes the fuel load variable as input to the DNN-fire models. In the coupled model, the DNN-Fire-GFED model would lead to results that are inconsistent with the feature space that was initially used to train the DNN-Fire model. Ideally, the authors should do a sensitivity analysis in the coupled DNN-Fire-GFED and ESM models to see if the results are still consistent and reliable. If this is not feasible, the authors should at least discuss how they would address such inconsistencies. I assume that only a joint optimization of fire and fuel loads/biomass in the coupled model would solve this issue (Drüke et al., 2019).

Response:

We agree that to fully couple DNN-Fire and E3SM land models is important and is our long-term objective for fire modeling. We will achieve this long-term goal with a stepwise approach. This study is the first step to develop and tune the 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, which we will pursue in future work. We appreciate the suggestion on joint optimization, and will explore the effectiveness of such an optimization strategy in the future.

In the revised manuscript, we add a paragraph to discuss the goal of this study and future work on fully coupled E3SM+DNN-Fire model (Line 356-359): "This study focuses on design, development, and parameterization of the DNN fire model within the E3SM model interface. In this way the DNN model can be readily coupled in the future and iteratively simulate climate, ecosystem fuel conditions, and fire dynamics. This study is an important step towards fully coupling E3SM and the DNN-Fire models in the future. "

2 Training and testing

The authors trained a DNN model for each GFED region. Training the model for different regions is an unfair approach in comparison to process-based fire models as these models are truly global models, maybe with a PFT-dependent parametrization. Hence the authors should provide a good reasoning why they trained the model per GFED region. In addition, it does make sense at all that a fire model is parametrised per GFED region for an application in an Earth system model. As Earth system models are applied to assess future changes, a parametrisation per region will fast lead to useless results. For example, if climate and vegetation conditions change in future, which regional model should be applied in a certain region? Fire should be only simulated as a response to climate, vegetation and socioeconomic conditions. If regional parametrisation is necessary, the parameters should be based on vegetation or socioeconomic conditions.

Response:

The choice of 14 fire regions was based on the historical convention from the Global Fire Emissions Database (GFED) studies. The 14 GFED regions were chosen based on clustering of fire behavior, background climatology, and vegetation types. The GFED regions also consider the suitability for comparison with other wildfire studies e.g., atmospheric tracer inversion studies (van der Werf 2006).

We appreciate the reviewer's suggestion to parameterize the DNN-Fire model based on vegetation types. We thus developed an alternative PFT-based parameterization strategy for the DNN-Fire model. We found that both PFT-based and GFED-based parameterization were equally accurate in terms of surrogating the original E3SM model and capturing the large-scale dynamics after calibration (Figure S7,S8).

Therefore, in the revised manuscript we added a paragraph to discuss the alternative PFT-based parameterization strategy (Line 392-398): "Our GFED region-based parameterization strategy relied on the combination of climate and biome types, while an alternative parameterization strategy for DNN-Fire model could be based on plant functional type distributions. Based on our analysis, the PFT-based DNN-Fire model had similar performance compared with the GFED-based model (Figure S7, S8). Since the current version of the E3SM land model does not allow PFT changes driven by climate, both GFED-based and PFT-based models may not fully capture the changes of fire dynamics due to longer-time scale fire regimes changes.”

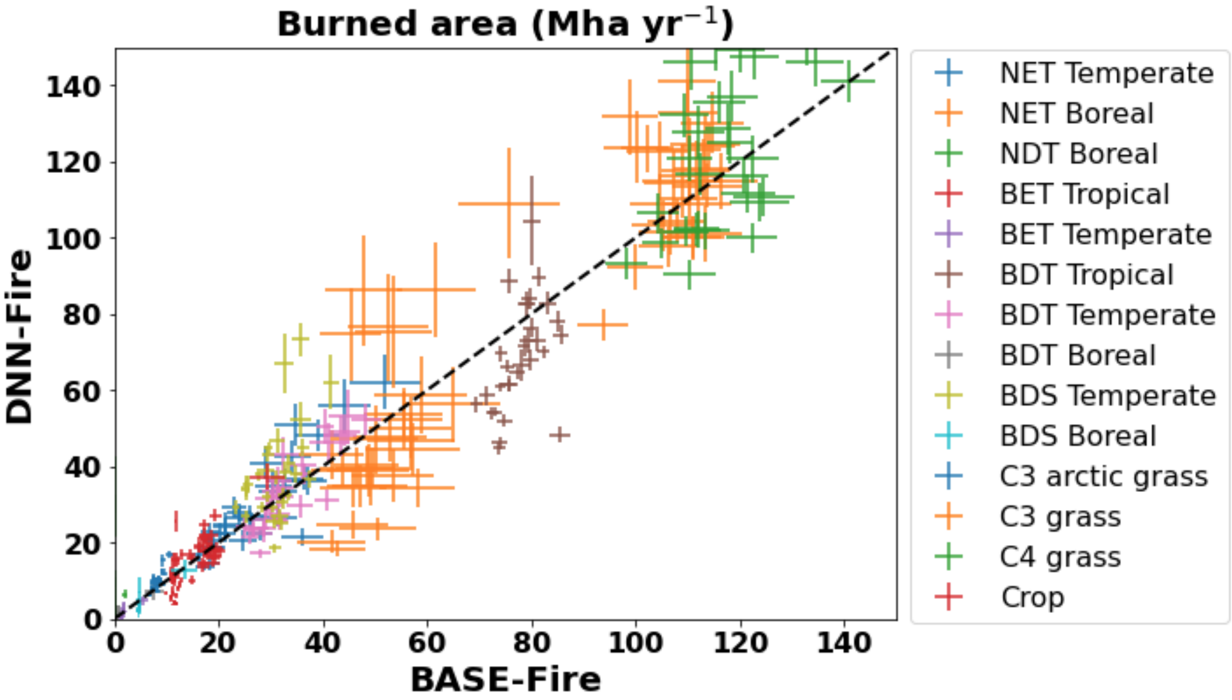

Figure S7. *The performance of the Deep Neural Network wildfire model (DNN-Fire), compared with the original ELMv1 process-based wildfire model (BASE-Fire) aggregated over 14 plant functional types between years 2001 and 2010.*

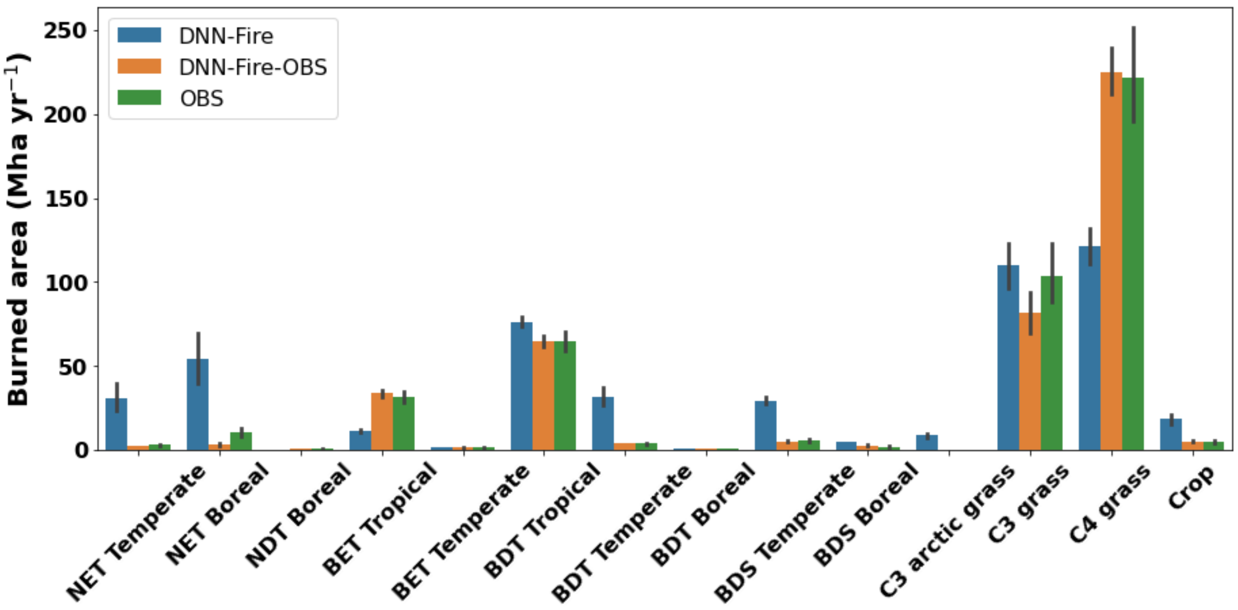

Figure S8. *A comparison of wildfire burned area among Deep Neural Network wildfire model (DNN-Fire), Deep Neural Network wildfire model fine-tuned with observed burned area (DNN-Fire-OBS), and observations for 14 plant functional types.*

The monthly burned area data from all grid cells in each regions was splitted randomely in 80% training data and 20% for testing. This is one of the simplest tests as the underlying conditions and statistical distribution of both samples is the same. However, in the context of an Earth system model, we expect non-stationary conditions and hence the model should be tested how well it can predict into 1) different regions, 2) different time periods (was done but the conditions in the two time periods are very similar), and 3) to different environmental conditions (Klemeš, 1986).

Response:

To address this comment and to maximally train and test the model across different fire conditions, we applied a "stratified random sample method" [Wang 2012] in the revised manuscript. The burned areas over all gridcells were first divided into three subgroups or "strata" based on the magnitude of the burn (low burn 0-33 percentile, medium burn 34-66 percentile, high burn 67-100 percentile). Then the gridcells were randomly sampled, but with the constraint that samples were drawn from each strata according to the ratios of samples within each strata. In this case, gridcells with different percentage burns (e.g., highly burned gridcells) were more likely divided into different datasets of training and testing, compared with the straightforward random sample method.

In the revised manuscript, we add a paragraph to describe and discuss the stratified random sampling approach (Line 199-205):

"Furthermore, the random sampling was stratified to reduce the risk of sampling, e.g., adjacent high fire gridcells. All gridcells were first divided into three "strata": low burn (0-33% percentile), median burn (33%-66% percentile), and high burn (67-100% percentile) gridcells based on the burn magnitude. The stratified random sample assured the sampled gridcells for training and testing had the same ratios of low, medium, and high burn, thus eliminating potential sampling bias from spatial autocorrelation [Wang et al., 2012]."

In order to assess the representativeness of the year chosen for training and testing, we trained and evaluated model performance with different years of test datasets 1) 2001-2002, 2) 2003-2004, 3) 2005-2006, 4) 2007-2008, 5) 2009-2010. The rests were used as training datasets, resulting in five different models, each trained by 8 years of data; and tested with the remaining 2 years of data. We found that the selection of training or testing years did not significantly change the model performance (Figure S1).

In the revised manuscript, we add a section to describe and discuss the impacts of selected year of test datasets on model performance (Line 205-211):

"In addition to random sampling, we also investigated the impacts of data choice on the model performance by sampling the testing datasets within specific years (e.g., 2001-2002, 2003-2004, 2005-2006, 2007-2008, 2009-2010) and using the rest of the years for training. We found insignificant differences among the models (Figure S1) indicating the choice of training and testing data years were not impactful. Therefore, we will discuss the results using the stratified random sampling approach throughout the paper."

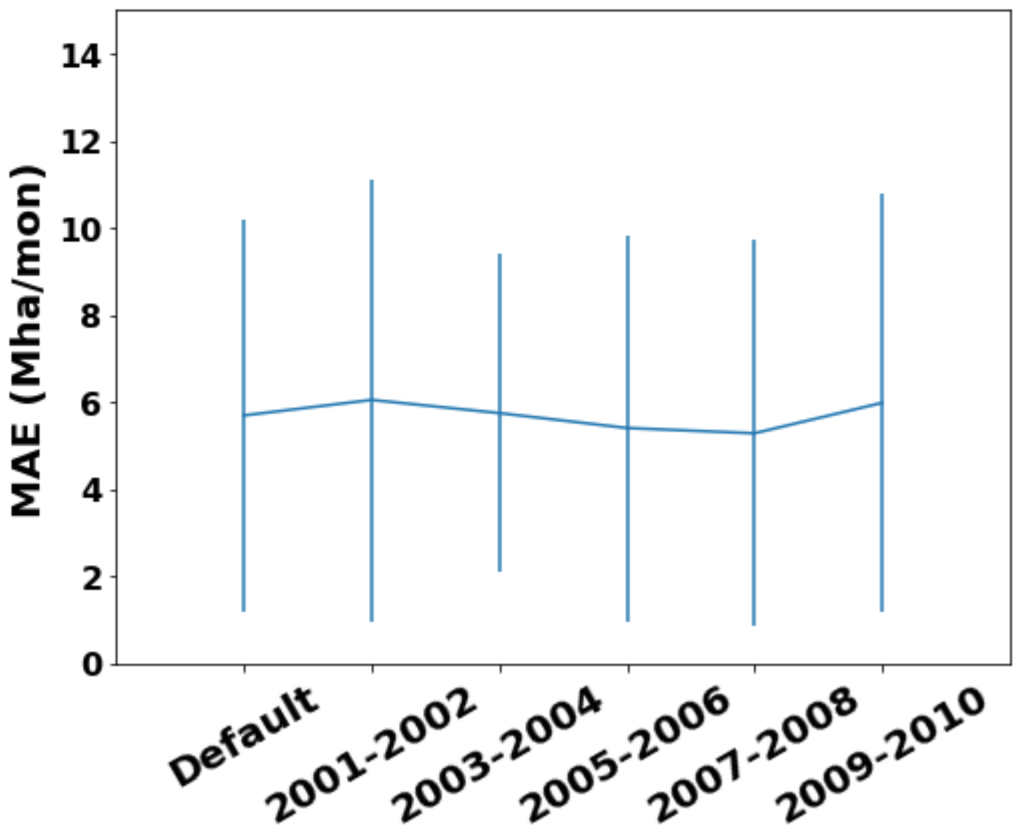

***Figure S1**. Model performance evaluated with testing datasets of default (20% randomly selected samples), or fixed to 2001-2002 period, 2003-2004 period, 2005-2006 period, 2007-2008 period, and 2009-2010 periods (the rest of the dataset was used as a training dataset.).*

3 Input data

Most of the input data for the DNN model comes from climate, land use or socioeconomic datasets. Information on fuel loads, fuel wetness and temperature, however, was taken from ELMv1 model simulations. I wonder about how good are these simulated variables in comparison with independent (e.g. Earth observation) data. For example, any biases in simulated biomass will directly affect the simulated burned area. Please compare the simulated biomass and soil moisture with useful datasets. Alternatively, a residual analysis would be also useful to see if any errors in simulated burned area rea related to errors in the simulated input.

Response:

To evaluate impacts of E3SM model simulated biomass and soil moisture on DNN-Fire model predictions, we drove the DNN model with the NOAA NCEP-NCAR-CDAS-1 (Kalnay 1996) soil moisture product. We found that soil moisture is not a significant source of DNN-Fire model uncertainty (Figure S5); in contrast, surface climate forcings overall were more impactful on *burned area simulations. For example, in the three largest fire regions (SHSA, NHAF, SHAF), dominant biases came from climate forcing rather than soil moisture (Figure S5).*

It is difficult to evaluate the bias from surface biomass within the DNN-Fire model, since continuous observed biomass data is not available. Thus, we directly compare the E3SM simulated long-term dynamics of vegetation carbon with present-day estimates (Figure S6). We found that E3SM reasonably captures the vegetation biomass stock.

In the revised manuscript, we add a paragraph to uncertainty from climate forcings, soil moisture, and fuel load (Line 360-379): "The DNN model uncertainty was subject to the accuracy of climate forcings as well as other physical driving variables simulated by the physical wildfire model (ELMv1). For example, in addition to the default GSWP3 climate forcings dataset used in the study, CRU-JRA [Onogi et al., 2007] and NCEP-DOE2 [Kanamitsu et al., 2002] reanalysis forcings were also widely used and potentially different from GSWP3 forcings. ELMv1 used climate forcing (e.g., temperature, precipitation, wind speed, relative humidity) to simulate soil temperature, soil moisture, fuel load and so on. These simulated variables served as inputs for the DNN model and could also result in prediction uncertainty. It was challenging to eliminate the forcing uncertainties in this work, but we could at least evaluate the magnitude of these uncertainties. We ran the DNN-Fire-OBS model with alternative forcings of CRU-JRA, NCEP-DOE2, and CDAS soil moisture from 2001 to 2010 and compared the results with DNN-Fire-OBS driven by default inputs (GSWP3 climate and ELMv1 simulated soil moisture) (Figure S5). The results showed relatively larger uncertainties from climate forcing than that from soil moisture forcing particularly over the major fire regions (e.g., SHSA, SHAF, and NHAF). For fuel load, although no transient dataset of global living biomass existed yet, we directly compared the ELM model simulated biomass with the global estimate (GEOCARBON ~ 455 Pg C). We found that the modeled present-day biomass continuously increased from 425 to 470 Pg C and compared reasonably well with the global benchmark. Future work will focus on evaluating the uncertainties from dead fuel load and fuel temperature variables."

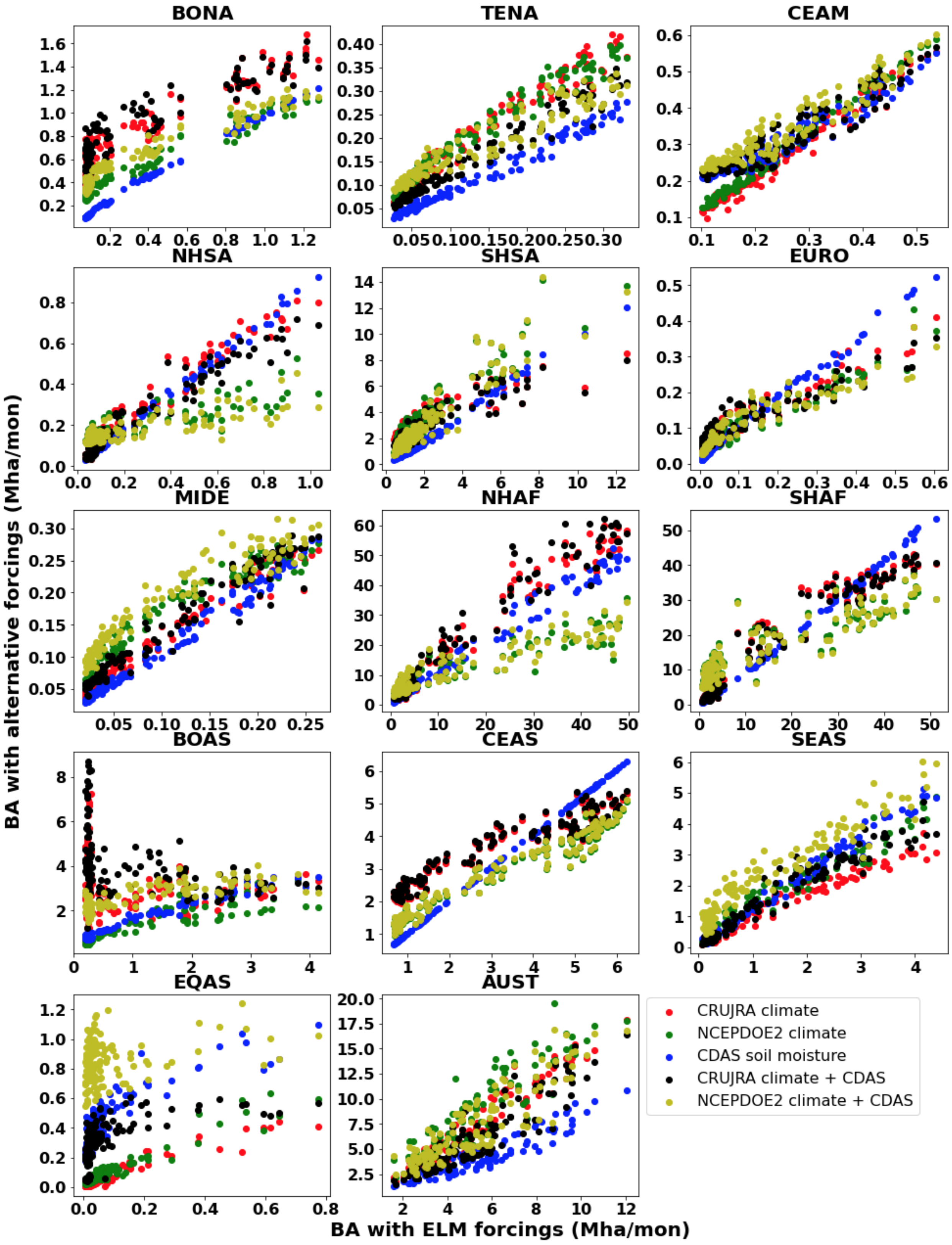

Figure S5. *Sensitivity of modeled burned area (2001-2010 long-term averaged) to climate forcings (including temperature, precipitation, wind speed, relative humidity) and soil moisture.*

X-axis is burned area simulated by the default model using GSWP3 climate forcing and ELMv1 simulated soil moisture. Y-axis is models with alternative climate forcing (CRUJRA, NCEPDOE2) and soil moisture (NCEP CDAS soil moisture) products.

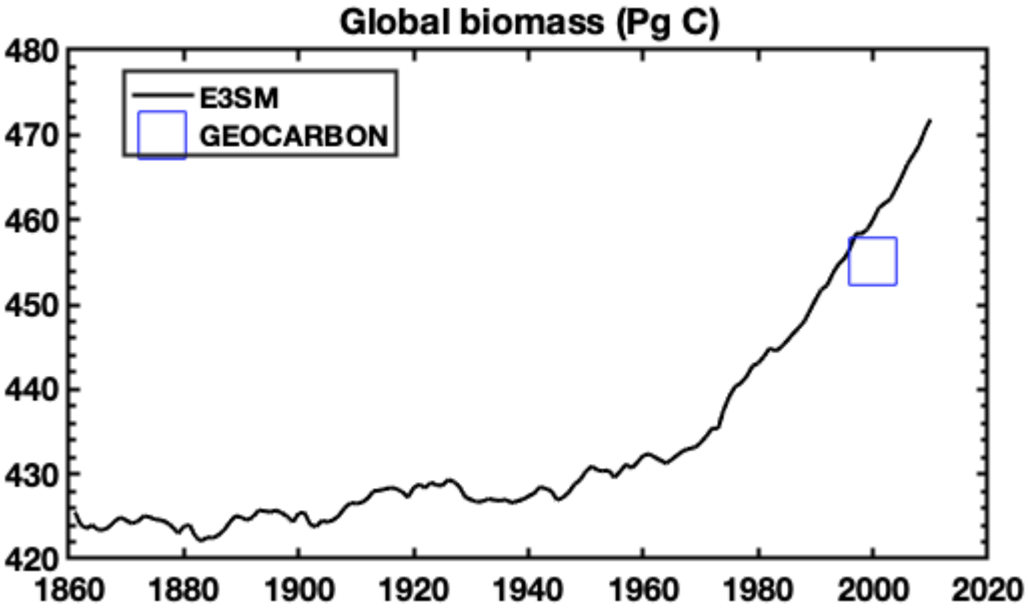

Figure S6. *E3SM simulated global vegetation biomass [425-472 PgC] and observational based estimate of present-day living biomass (455 PgC GEOCARBON).*

Can you please demonstrate that the tree cover from the LUH2 dataset is consistent with the simulated biomass. Are there any areas where the simulated biomass does not correspond to tree cover?

Response:

LUH2 land cover change time series are prescribed as forcing variables within E3SM including tree cover, therefore consistency between E3SM and LUH2 is imposed in the model.

Specific comments

L 26-27: From this statement it is not clear if the DNN is implemented as part of the E3SM or if it is independent of the ESM and just returns the same output. Please clarify

Response:

In the revised manuscript, we have clarified this point with "with the Energy Exascale Earth System Model (E3SM) interface", as described above.

L 30-31: It is not clear what the R2 means. Is it the R2 between the observed and predicted global annual total burned area in 2001 and 2015?

Response:

In the revised manuscript, we have clarified this point with: "The surrogate wildfire model successfully captured the observed monthly regional burned area during validation period 2011 to 2015 (coefficient of determination, *R^2 = 0.93)"*

L 41: The statement should be updated with newer estimates, e.g. by (Lasslop et al., 2020)

Response:

In the revised manuscript, we have updated the sentence to read: "global forests would double if fire were eliminated [Bond et al., 2005; Lasslop et al., 2020]"

L 78-93: You should clarify the scale of wildfire models. Fire behaviour models aim to model the spread and intensity of individual fires and are widely used in fire management. Fire models as parts of global vegetation or Earth system models have a different purpose. I assume that you are mainly addressing the second group of models, so please clarify it. Here you should specify that the first group focus mostly on predicting large scale regional fire dynamics, whereas the second group focus more on predicting fire in individual grid cells.

Response:

*In the revised manuscript, we have clarified these points as (line 93-100): "*Historically, data-driven models were often used for fire behavior modeling to predict ignition, spread, duration, and extinction of individual fires [*Finney*, 1998; *Radke et al.*, 2019] at fine spatial and temporal scales. This group of models are more relevant to operational fire research. In contrast, process-based wildfire models used in global vegetation models or earth system land models focus on gridcell aggregated fire burned area dynamics that are more relevant to analyses of large-scale patterns and climate-scale predictions [*Fang Li et al.*, 2019; *Rabin et al.*, 2017]. This study particularly focuses on the second category of wildfire models.*"*

Chapter 2.2: The text might be easier to understand if you draw the network structure as a figure including all input variables, the hidden layers, neurons and output.

Response:

We have updated Figure 1 to reflect the input variables and structure of DNN models.

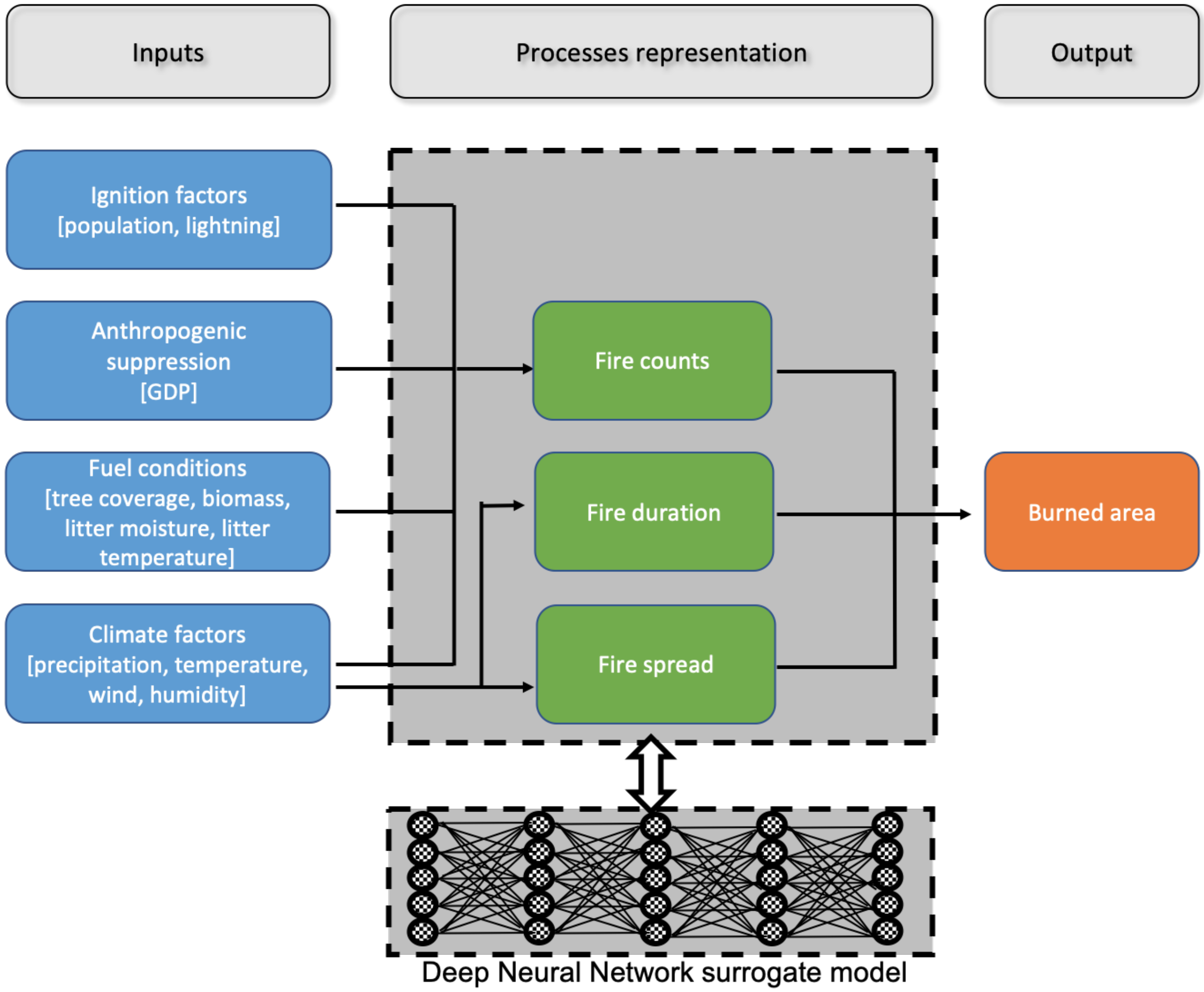
Inputs
Processes representation
Output
Ignition factors [population, lightning]
Anthropogenic suppression [GDP]
Fuel conditions [tree coverage, biomass, litter moisture, litter temperature]
Climate factors [precipitation, temperature, wind, humidity]
Fire counts
Fire duration
Fire spread
Burned area
Deep Neural Network surrogate model

Figure 1. *Schematic representation of the ELMv1 process-based BASE-Fire model and the components to be surrogated with the Deep Neural Network (DNN) model (dark grey).*

L 163-171: The description of the training of DNN-fire-GFED is not completely clear. From the text it reads that only the weights were readjusted by using observed GFED data. Does that mean that original bias parameters from DNN-Fire-BASE were kept? Is there any reasoning?

Response:

We adopted the standard transfer learning approach [Do et al., 2005] that, first, pre-trained the DNN-fire model with E3SM outputs to generate reasonable baseline values for weight parameter, and second, using the pre-trained weight parameters as initial values and then fine-tune the weight parameters using observations.

L 180: “spunup”

Response:

“spunup” was corrected in the revised manuscript.

L 197-201: The readability would be improved if each equation is in a new line and not within the text line.

Response:

Equations 9-11 are updated in the revised manuscript.

L 244: Should this be Figure 7?

Response:

Corrected.

L 273-275: Yes, but not many process-based fire models have been really calibrated. It would be good to provide examples in the text where this has been done.

Response:

We have removed the ambiguous statement.

L 276-277: The statement is not really valid as you do not calibrate the parameters of the process-based model but of the DNN-based model.

Response:

We have removed the ambiguous statement.

L 332-334: I do not understand this sentence because you previously wrote that you were training models for different regions and not a global model. Please clarify.

Response:

ELM process-based fire model (not DNN surrogate model) has a unified representation for global wildfire dynamics.

Table 1: It would be good to combine the columns data source and reference in one column. Otherwise it seems odd because population density and GDP do not have a data source.

Response:

Data source and reference columns are combined in Table 1.

Figure 1: check “burn” area

Response:

Corrected.

Figures 3, 5, 6: I recommend to combine these figures in one figure (with 4 columns per region) in order to directly compare the experiments in one plot. In addition, it would be good to also draw in a same way boxplots or violin plots of monthly burned area in order to check if the different experiments capture the statistical distribution of fire.

Response:

We appreciate the recommendation. We have combined Figure 3, 5, and 6 into one figure (Figure 3).

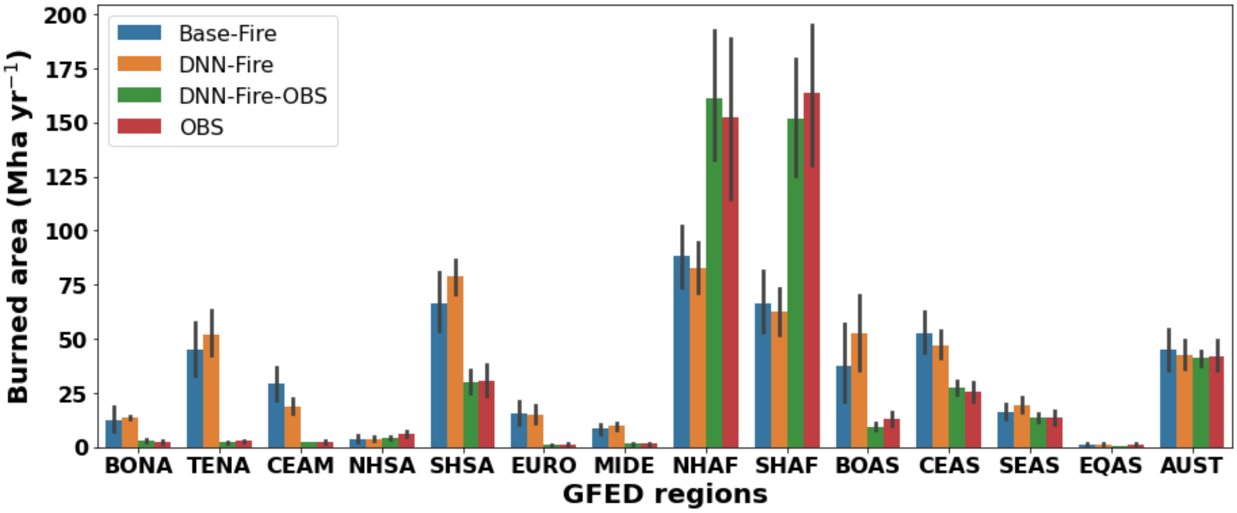

Figure 3. *A comparison of wildfire burned area between estimates from the ELMv1 process-based model (BASE-Fire), Deep Neural Network wildfire model (DNN-Fire), Deep Neural Network wildfire model fine-tuned with observed burned area (DNN-Fire-OBS), and observations over 14 GFED fire regions.*

Figure 4: This figure includes a lot of spatial aggregation. Can you draw a density scatter plot of the original monthly data in the used 1.9 x 2.5° resolution?

Response:

Thanks for the suggestions. We have added a density scatter plot in the supplementary material to demonstrate the performance of the surrogate model. The scatter plot showed that the majority of the BASE-Fire variability was captured by the DNN-Fire surrogate model (high density regions lie on 1:1 line).

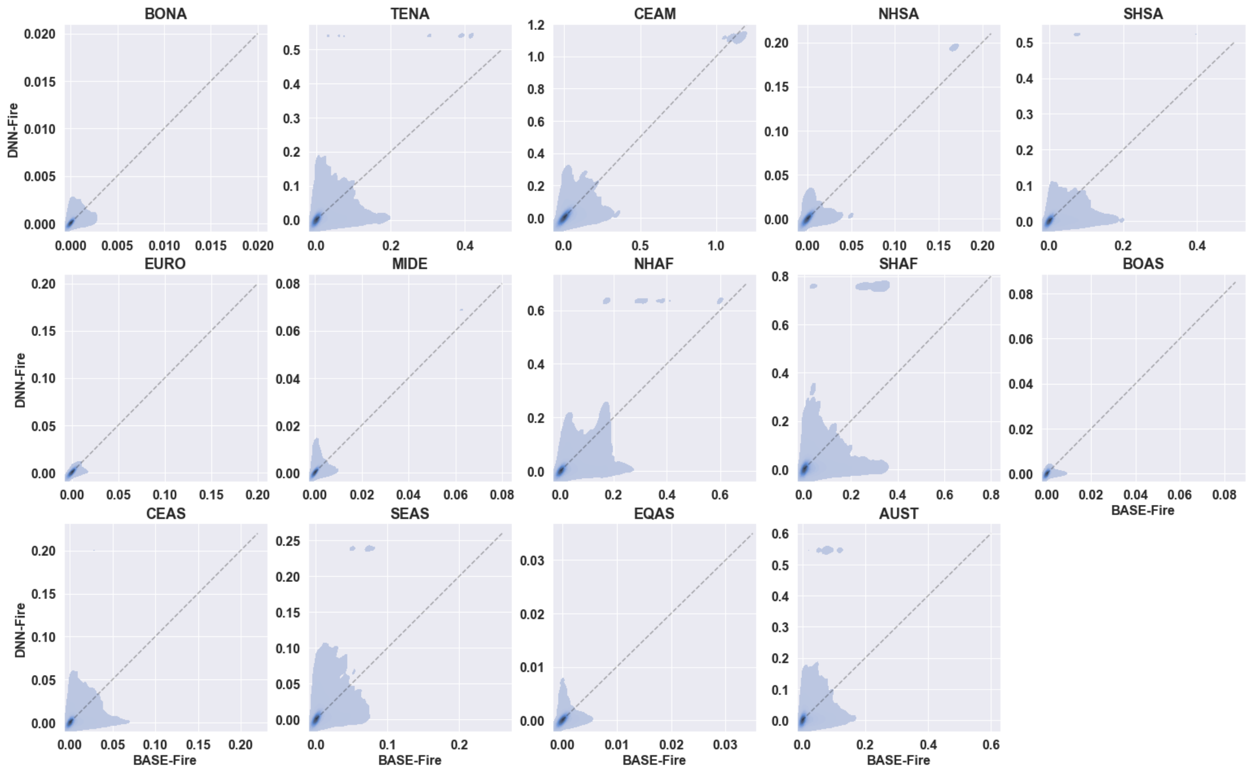

Figure S2. *Performance of surrogate model (DNN-Fire) compared with ELMv1 process-based model (BASE-Fire).*

Figure 7 b: Is this a global averaged seasonal cycle? How do the seasonal cycles look like in different GFED regions?

Response:

Figure 7b is the global average and therefore dominated by major GFED fire regions, i.e. NHAF, SHAF, SHSA. For each different GFED region, we added a new figure in supplementary material to illustrate the seasonal cycles of modeled and observed burned area. Overall, the DNN-Fire-OBS did a reasonably good job in capturing the seasonal dynamics of the burned area. Model biases were found for some specific months of the year. For example, DNN-Fire-OBS missed the decline of burned area in June over TENA and the relatively low burned area in March and April over AUST.

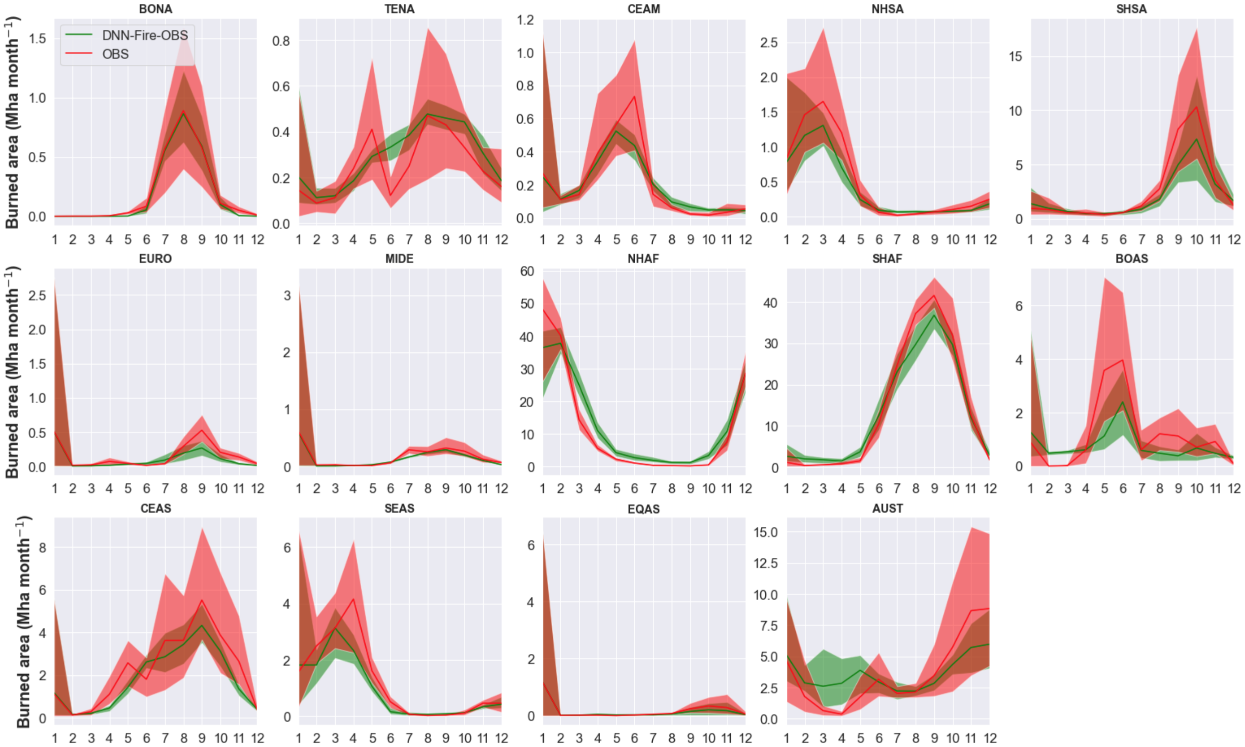

Figure S3. *Seasonal cycles of fine-tuned Deep Neural Network wildfire model (DNN-Fire-OBS) and observations over 14 GFED fire regions.*

Reference

Do, C.B. and Ng, A.Y., 2005. Transfer learning for text classification. Advances in neural information processing systems, 18, pp.299-306.