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_CCILT11	*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}

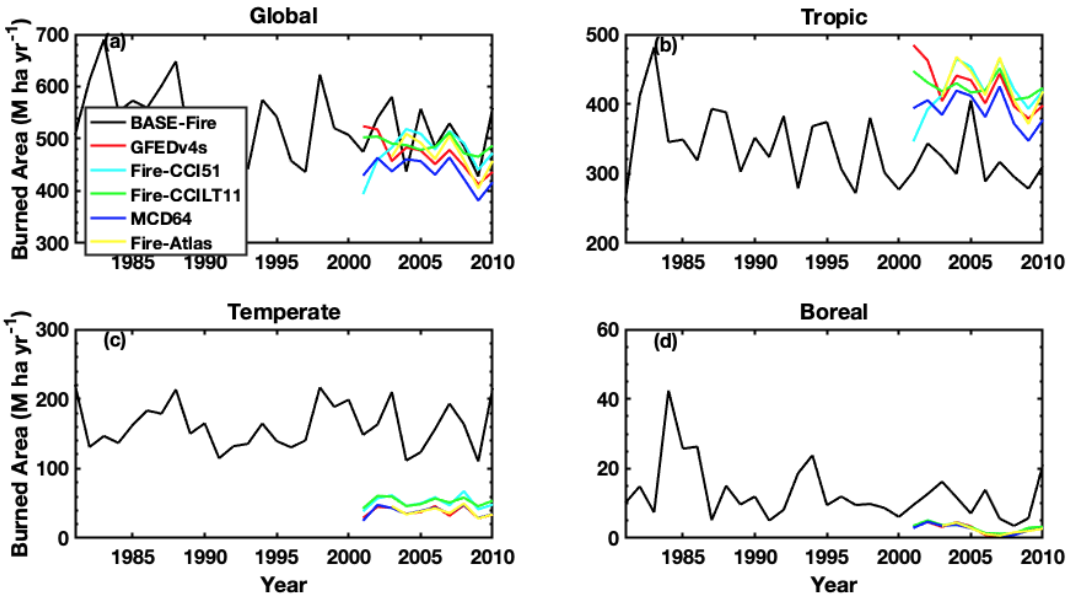

Figure 2. *BASE-Fire simulated and burned area datasets of GFEDv4s, Fire-CCI51, Fire-CCILT11, MCD64, Fire-Atlas. (a) Global scale; (b) Tropical (S23.5° -N23.5°); (c) Temperate (N23.5° - N 67.5°); and (d) Boreal (north of N 67.5°) regions.*

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."

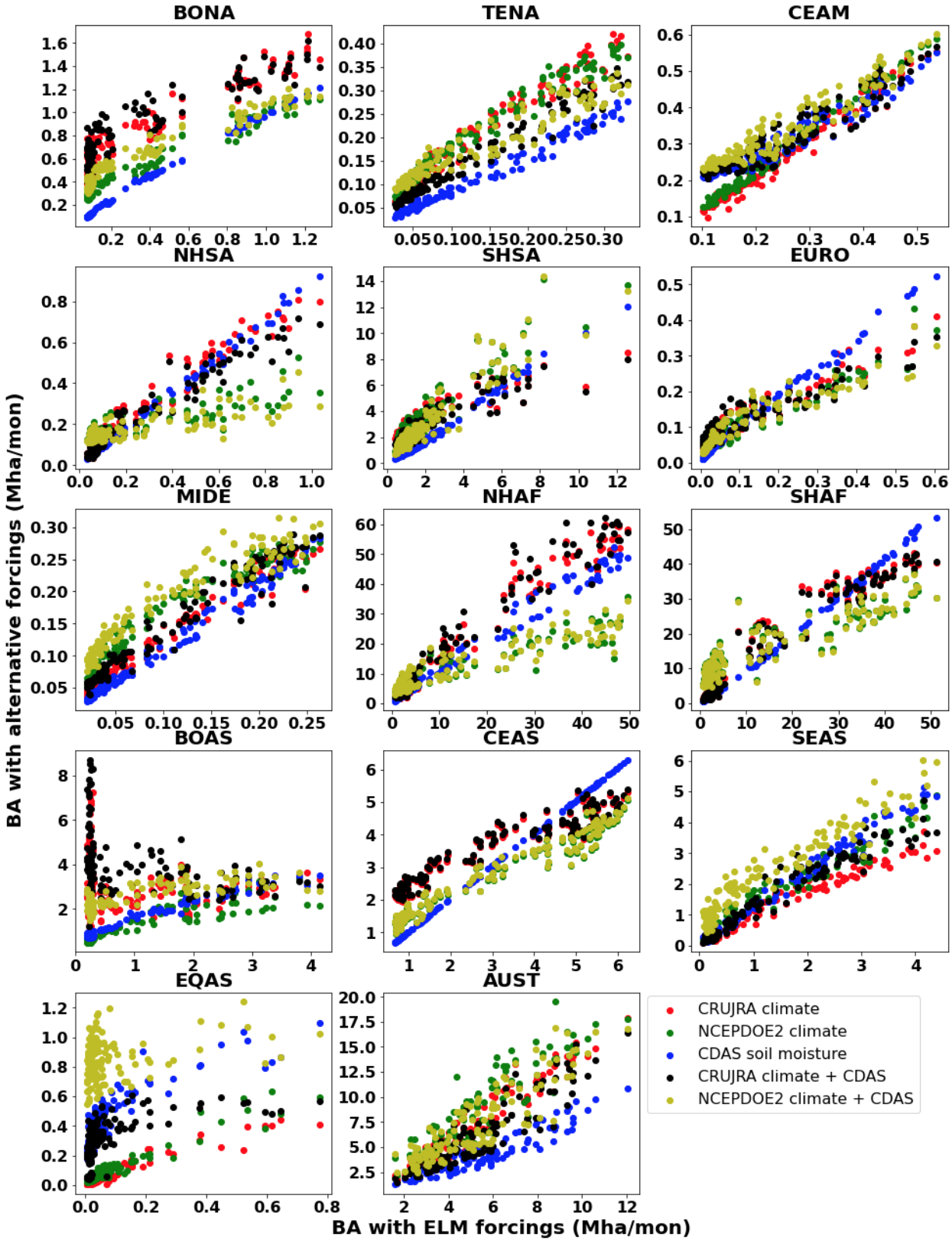

Figure S3. *Sensitivity of modeled burned area (2001-2010 long-term averaged) to climate forcings (including temperature, precipitation, wind speed, relative humidity) and*

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."

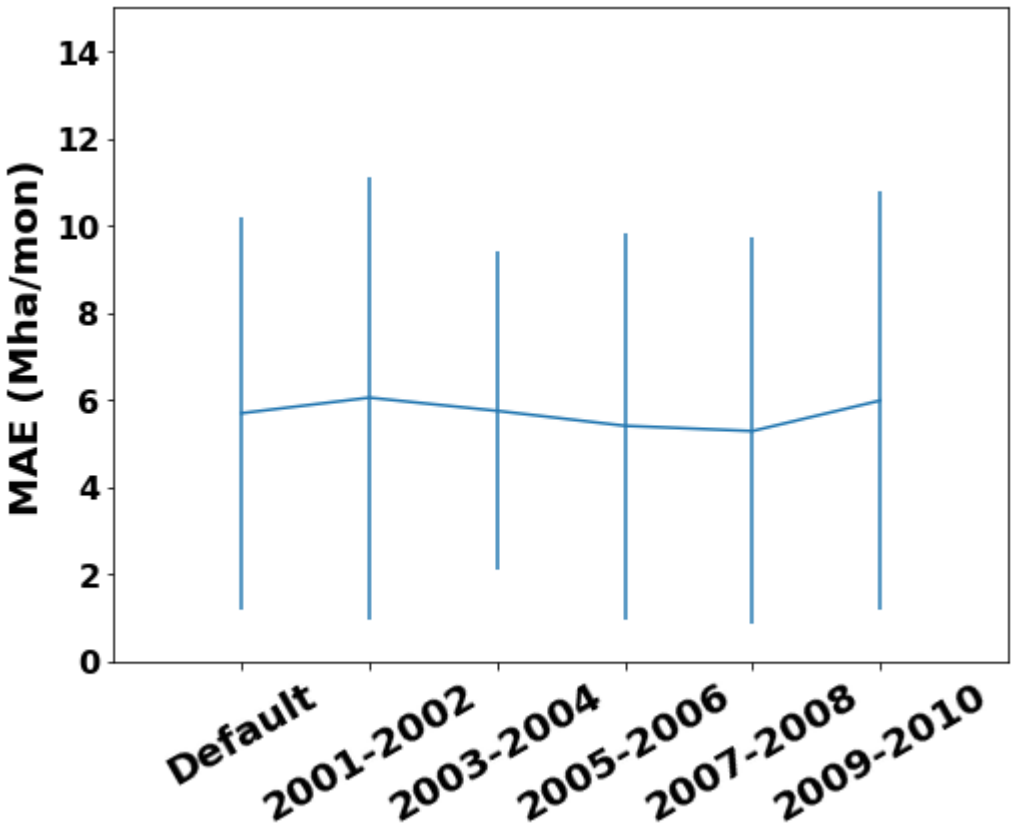

Figure S1*. Model performance evaluated with testing datasets of default (20% randomly selected samples), or fixed to 2001-2002 period, 2003-2004 period, 2005-2006 period, 2007-2008 period, and 2009-2010 periods (the rest of the dataset was used as a training dataset.).*

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} ."

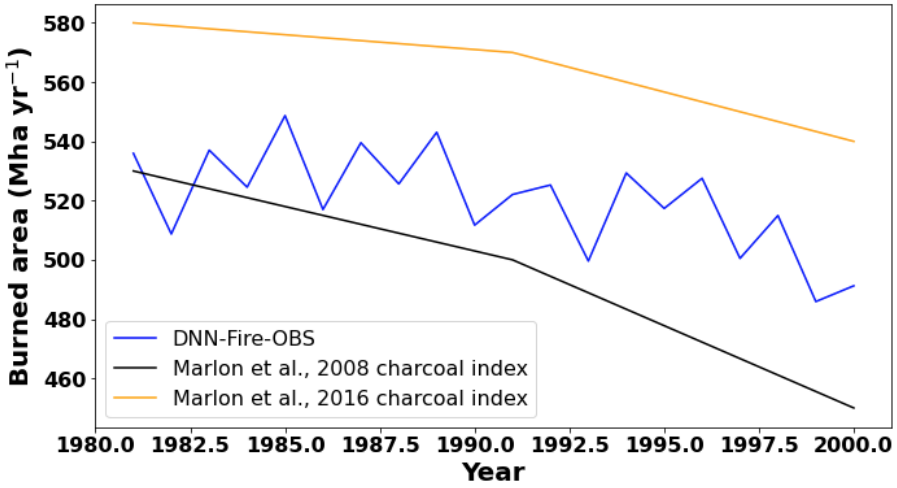

Figure S2. *Comparison of DNN-Fire-OBS model simulated global burned area during 1981-1999 with two charcoal index inferred burned area.*

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."

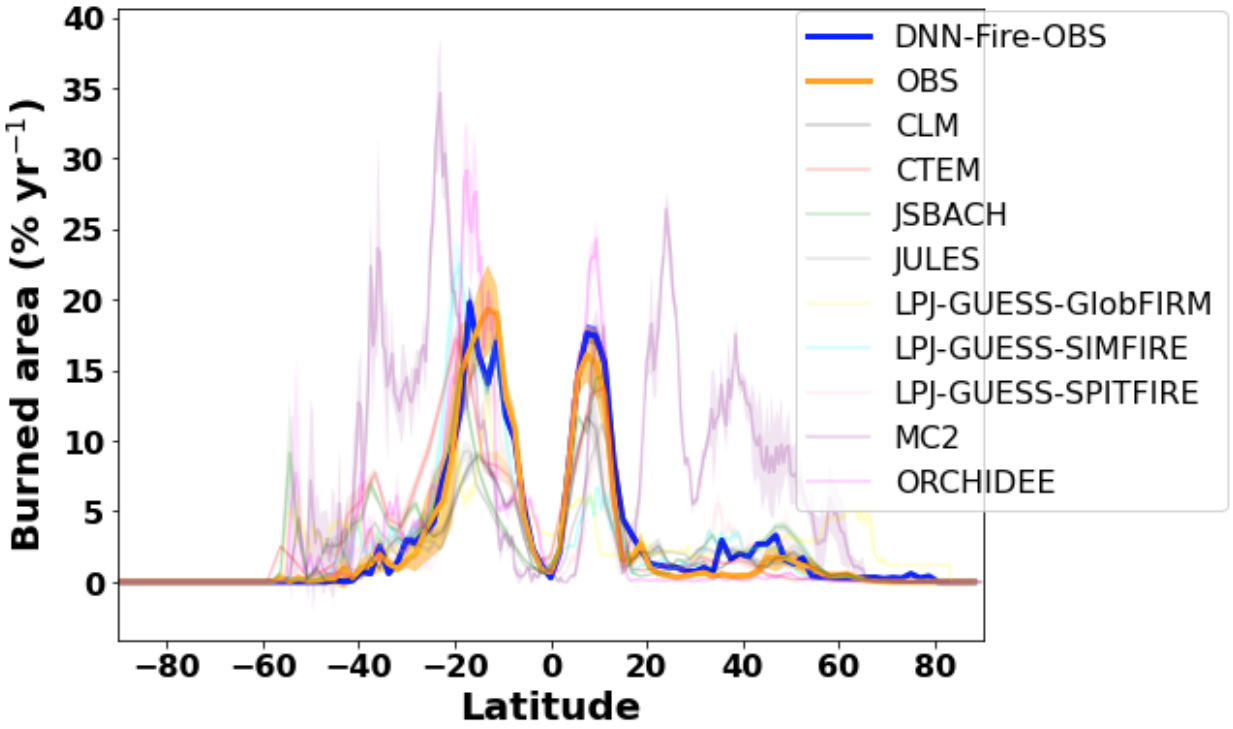

Figure 9. *Prognostic simulation of annual wildfire burned area (2011-2015) with the Deep Neural Network wildfire model fine-tuned with observations (DNN-Fire-OBS) compared with observations and nine FireMIP models outputs.*

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.

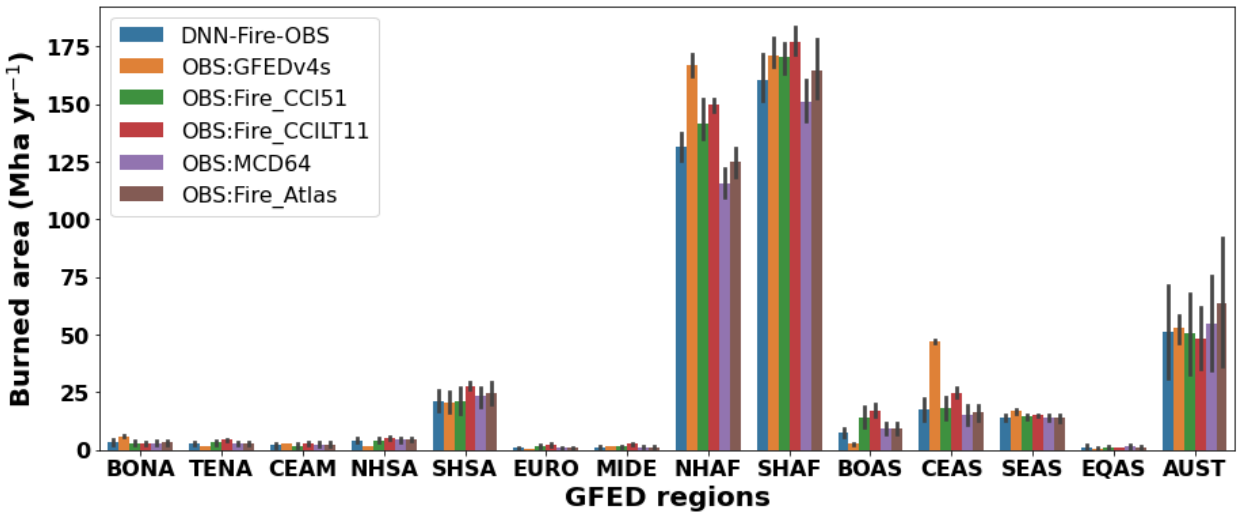

***Figure 8**. Prognostic simulation of annual wildfire burned area (2011-2015) with the Deep Neural Network wildfire model fine-tuned with observations (DNN-Fire-OBS) compared with five observational datasets.*

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.