We sincerely appreciate the referees for their valuable and insightful comments on our manuscript. The feedback is instrumental in enhancing the quality and clarity of our research. These comments are not only valuable but also serve as a critical resource for improving various aspects of our article, including methodology, data interpretation, and overall presentation. We have taken each comment seriously and conducted a thorough review of our manuscript to ensure that we comprehensively address all concerns raised by the referees.

This response document provides a detailed account of the changes implemented in relation to each specific comment from the referee. For ease of reference, referee comments are presented in black, while author responses are highlighted in blue, with modifications to the manuscript in italic font. The line numbers correspond to the clean version of the revision.

General comments

This manuscript builds on previous work that used climate forcing observations and vegetation model–derived vegetation outputs to build a fire model over the continental U.S. (CONUS) using the XGBoost machine learning algorithm. Here, the authors couple that fire model back into the ELM land and vegetation model, resulting in marked improvements relative to the built-in, process-based ELM fire model in terms of total burned area, its seasonal timing, and its interannual variability. There is (as expected) some decrease in performance relative to the uncoupled ML fire model, but not much. The authors also compare their ELM simulations with other process-based fire models in the FireMIP experiments. The manuscript is mostly well-structured, the figures are easy to understand, and the writing is for the most part clean and clear.

Process-based fire models are notoriously complicated and uncertain, so I am quite interested in the potential of machine learning to supplement, complement, or even replace them. However, I have serious concerns about the usefulness of the particular model system described here. I also have various less-severe but still-important concerns related to methodological and analytical issues.

To some extent these can be addressed by expanding the Discussion and adding subsections for organization. The authors should reduce the amount of space in the Discussion dedicated to reiterating already-stated results, instead only re-presenting results as needed to support new assertions. However, my fundamental concern about the usefulness of the model system presented here will require a fair amount of additional work. I thus recommend this paper be reconsidered after major revisions.

Thank you for your thorough review and valuable insights. We agree that the discussion section would benefit from additional subsections to better organize key points and reduce repeated results. We will reorganize this section, highlighting methodological implications and comparison insights while minimizing redundancy. Additionally, we fully agree with your perspective on the roles of process-based models and machine learning approaches. Process-based models provide critical insights into mechanistic processes, while machine learning approaches offer opportunities to capture complex patterns that may be challenging to model directly. We believe combining these methods can enhance predictive capabilities and address uncertainties effectively. We have reorganized the Discussion section to avoid repeating contents and extend the discussions.

Specific comments

Utility or "fitness for purpose" of ML-based fire model

First, I want to outline what this concern is not about. There is a long-standing philosophical question of whether an empirical model can ever be trusted outside the time period and/or environmental conditions in which it was trained. Climate change and socioeconomic developments are expected to introduce never-before-seen combinations of environmental conditions and human behavior, requiring extrapolation. Some people considering this issue conclude that only process-based models are useful. However, I'm *not* making that argument right now—my mindset is that process-based models are also imperfect, so empirical and ML models can be useful as well.

My concern is more that methodological issues in this paper make it so that I'm not sure of the usefulness of *this particular* model system. There are two main reasons for this.

First, there is not actually one "offline-XGB" fire model, but rather twenty—one for each year. The authors did this in an attempt to avoid "overfitting." Their use of that word doesn't fit with how I understand it, so I interpreted it as them avoiding training and testing their model on the same data. This is an important goal, but by choosing to do it this way, there is no single model presented that could be used for years outside 2001–2020. Instead, randomly excluding (e.g.) 20% of gridcell-years and building one model based on the rest would allow the construction of one canonical model that could be used for prognostic simulations. (In addition, it's unclear whether the authors mask low-fire gridcells only in training or also in simulations—if the latter, prognostic simulations would of course always predict zero burned area there.) This means that the ELM+offline-XGB model isn't actually useful as a predictive tool, contrary to the authors' assertions (e.g., P14 L18–19, P16–17 L72–76).

We fully understand your concern. The concept of splitting data by years and building a separate model for each year was to ensure all the predictions were not using data that the machine learning model had seen during its training process. However, we acknowledge that this approach oversights the usefulness of this model. Regarding masking low-fire grid cells, we only applied the mask during the training process.

In the revision, we have trained a canonical model using the random split approach, 80% for training and 20% for validation. The canonical model retains the performance in both offline and coupled models, and this single model will be useful in prognostic simulations. For more details, please see the responses to the specific comment.

So if the presented model isn't useful for prediction, can it help us understand anything about the drivers of present-day burned area and its trends (as the authors try with the detrended-temperature experiment)? Unfortunately, the answer to that question is also no, because it's trained on unreliable model outputs of vegetation biomass, composition, and dynamics. This is a real concern with highly-tuned models, including ML models, which can end up being "right for the wrong reason," using one process to compensate for another that's poorly-represented. While observational data are also imperfect, an ML model trained only on observations—especially one designed with explainability in mind— would be more trustworthy when it comes to examining the influence of different drivers. I think the authors could resolve this usefulness issue by building one canonical offline-XGB fire model, enabling its use in prognostic simulations. I don't really have a problem with this being trained on ELM-simulated vegetation data; yes, that will mean compensation of ELM's biases, but that can happen in pure process-based models anyway. However, it does mean that the detrended-temperature analysis should be removed from the paper, unless both the ML and process-based models' temperature responses are compared to a purely-observation-based analysis. Removing that analysis would be fine for me, as I find it somewhat extraneous.

We appreciate your understanding of training the ML model with ELM-simulated vegetation data and your insights about model error compensation. We acknowledge the limitation of attribution analysis using a highly-turned model. In the revision, the analysis and discussion on the detrended temperature have been removed.

Methodological questions

• Sect. 2.1.2:

This should be expanded to include a brief summary of how the "pretrained" model worked—enough for the reader to understand what "the large-scale patterns" are without having to consult Wang et al. (2021).

We have added the following texts to expand the introduction on the pretrained XGBoost model (Lines 115-123).

In this study, we adapted the XGBoost algorithm used by Wang et al (2021) to develop an offline ML fire model using variables directly provided by ELM at each grid cell. Wang et al. (2021) integrated large-scale meteorological patterns alongside local weather, land surface properties, and socioeconomic data to enhance the prediction of burned areas. The large-scale patterns were identified using singular value decomposition (SVD) to capture influential atmospheric conditions that develop over days to weeks and cumulatively impact the monthly burned area. The feature importance analysis in their study noted that while large-scale patterns improved prediction, however, they played a secondary role. Therefore, we exclude the large-scale patterns from predictors without significantly affecting the model accuracy. Hereafter the uncoupled XGBoost fire model is referred to as offline-XGB.

It's also a bit confusing to say you're using the "pretrained" model, but then you change how it works and retrain it for each year from 2001–2020. Was "pretrained" supposed to mean that it's being trained offline, i.e., before being coupled?

Thanks for raising this great point. In the revision, we modified this sentence to:

In this study, we adapted the XGBoost algorithm used by Wang et al (2021) to develop an offline ML fire model using variables directly provided by ELM at each grid cell.

And

Hereafter the uncoupled XGBoost fire model is referred to as offline-XGB.

Excluding low-burning gridcells: Is this just in training, or are they also masked in simulations? If the latter, then the ELM-BGC and FireMIP outputs should also be masked.

The low-burning grid cell mask is only applied in the training process. We have clarified this in the revision.

- Analysis of effect of rising temperature: Why is ML4Fire-XGB included but not offline-XGB? The latter is more what this paper is actually about.

In the original manuscript, we used ML4Fire-XGB in the rising temperature experiment to account for the temperature effect on vegetation growth and its consequence on the burned area. However, we agree with the reviewer's inspection of this analysis, and we have removed the relevant analysis and discussion.

- It's unclear what FireMIP outputs you used. P6 L43: In addition to Rabin et al. (2017), the FireMIP phase 1 burned area publication should also be cited. This might be Hantson et al. (2020, doi:10.5194/gmd-13-3299-2020), but the models chosen here aren't all present in that publication. Hopefully Rabin et al. (2017) describes the simulation protocol for the models whose output you've chosen to compare; if not, a different publication that includes the protocol should be cited.

Thank you for pointing out the confusion. In our manuscript, we used outputs from the latest FireMIP models, i.e., the FireMIP Phase II or the ISIMIP-Fire sector (ISIMIP3a) (https://protocol.isimip.org/#/ISIMIP3a/fire). The ISIMIP3a experimental design follows the protocol outlined in Rabin et al. (2017), with all models using a common set of climate and socioeconomic (land-use, GDP etc) data provided by ISIMIP3a. We opted for ISIMIP3a model outputs due to the updates in fire models implemented and a longer simulation period post-2000 in this phase. Although there is no specific protocol paper for ISIMIP3a, two publications have recently been made available (Burton et al., 2024; Park et al., 2024). At the time of our study, only four models had uploaded results, but we have now updated our analysis to include outputs from all seven ISIMIP3a models, as detailed in Burton et al. (2024).

Reference:

Burton, C., Lampe, S., Kelley, D. I., Thiery, W., Hantson, S., Christidis, N., Gudmundsson, L., Forrest, M., Burke, E., Chang, J., Huang, H., Ito, A., Kou-Giesbrecht, S., Lasslop, G., Li, W., Nieradzik, L., Li, F., Chen, Y., Randerson, J., Reyer, C. P. O., and Mengel, M.: Global burned area increasingly explained by climate change, Nat Clim Change, 10.1038/s41558-024-02140-w, 2024.

Park, C. Y., Takahashi, K., Fujimori, S., Jansakoo, T., Burton, C., Huang, H., Kou-Giesbrecht, S., Reyer, C. P. O., Mengel, M., Burke, E., Li, F., Hantson, S., Takakura, J., Lee, D. K., and Hasegawa, T.: Attributing human mortality from fire PM2.5 to climate change, Nat Clim Change, 10.1038/s41558-024-02149-1, 2024.

Inconsistency of comparisons

- This manuscript compares ML-based model performance against GFED5—were the process-based models parameterized against that? Probably not, because it's pretty new. GFED5 has 61% more burned area than GFED4s (which, the process-based fire models may have been calibrated against GFED4 or even 3). Although I'm not sure how much the increase was in CONUS. This should all be explored in the Discussion.

We appreciate the reviewer's concern, particularly as we developed one of the process-based models participating in ISIMIP3a, and GFED4s served as the reference dataset for global fire calibration in these models. For the ML model training, we chose GFED5 because it has been shown to better capture small fires compared to earlier datasets (Chen et al., 2023; Roteta et al., 2019), which often under-represent prevalent agricultural fires in the Central U.S. Additionally, GFED5 is now used as the reference dataset in the latest FireMIP/ISIMIP3a publication (e.g., Burton et al., 2024). While differences in magnitude exist between GFED5 and GFED4s in burned area estimates within CONUS, these datasets also share common features. The spatial correlations of GFED4s and FireCCI5.1 against GFED5 are over 0.66. In the revision, we have added a new figure (Figure 3, also see below) comparing GFED5, GFED4s, and FireCCI5.1. The process-based models indeed face challenges in accurately predicting burned areas over CONUS, even when evaluated against GFED4s or FireCCI5.1.

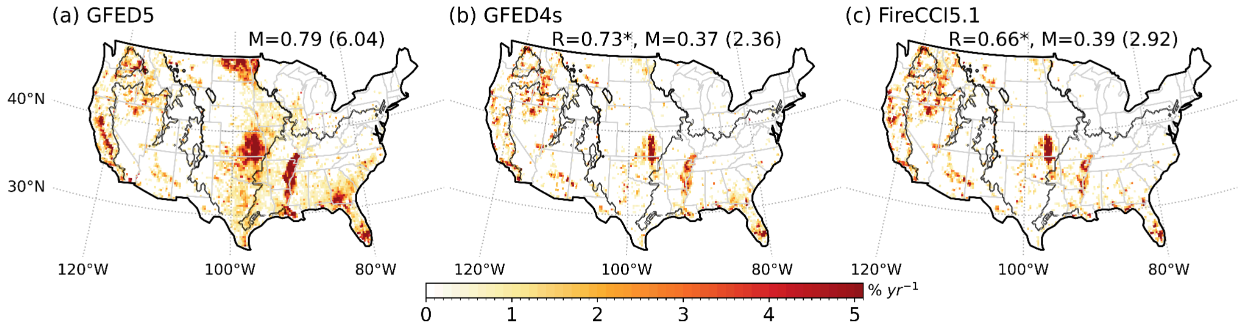


Figure 3: Observed burned area fraction (% yr-1). (a) GFED5 (2001-2019), (b) GFED4s (2001-2016), and (c) FireCCI5.1 (2001-2019). The numbers indicate the mean (M) burned area fraction and burned area (in Mha) in brackets for each dataset. The pattern correlation (R) against GFED5 is also shown, with an asterisk (*) denoting significance at the 0.01 level. Black contours outline the ecoregions.

Reference

Chen, Y. et al. Multi-decadal trends and variability in burned area from the fifth version of the Global Fire Emissions Database (GFED5). Earth Syst Sci Data 15, 5227–5259 (2023).

Roteta, E., Bastarrika, A., Padilla, M., Storm, T. & Chuvieco, E. Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa. Remote Sens Environ 222, 1–17 (2019).

- I’m not sure exactly what FireMIP simulations you used, but they almost certainly used different climate, lightning, population density, and/or GDP inputs from the ELM simulations here (as well as the uncoupled ML4Fire-XGB training and usage).

This study uses model simulations from the ISIMIP3a (FireMIP phase II). In our ELM-BGC and ELM2.1-XGBfire1.0 (the coupled ELM and XGB fire model) simulations, we adopted the same lightning, CO_2, population density, and GDP data used in ISIMIP3a, with the exception of the climate forcing data. To focus on fires in CONUS, we applied the hourly NLDAS climate forcing at a spatial resolution of 0.25º, rather than the daily GSWP3-W5E5 forcing at 0.5º used in ISIMIP3a. This different reanalysis data source and differences in the spatial and temporal resolutions of the climate forcing could contribute to variations in burned area predictions.

Besides ISIMIP3a models, we also conducted ELM-BGC (with built-in process-based fire model) simulations driven by the same set of climate, lightning, and socioeconomic forcing data as used to drive the coupled model ELM2.1-XGBfire1.0. The results show that the burned area simulation in ELM-BGC remains unsatisfactory, indicating that changes in climate forcing alone do not account for all limitations in burned area simulations in process-based models (at least for ELM-BGC). We have added the following discussion to the revised manuscript to clarify this point (Lines 360-363).

The ISIMIP3a models were driven by daily GSWP3-W5E5 forcings at a 0.5º spatial resolution. Differences in forcing data could lead to variations in burned area predictions. However, since both ELM-BGC and ELM2.1-XGBfire1.0 are driven by the same set of forcings, this suggests that limitations in physical processes may significantly hinder the performance of process-based models.

- After reading the Results, the fact that the process-based models all overestimate burned area in CONUS despite (probably) having been calibrated against the (probably) lower CONUS burned area observations suggests that the process-based models are extra wrong!

The reviewer’s observation is correct. The estimation over the CONUS in GFED5 is 114% larger than GFED4s (Figure 3 in the revision). As we discussed in the manuscript, the process-based models often focus on the globe or fire-prone regions such as African savannas, where fire regimes can be distinct from the CONUS. We believe process-based model performance over the CONUS can be improved with parameter calibration and an advanced understanding of the missing physics.

FireMIP models

- Why were only those four FireMIP models chosen? This question is especially important because, as you note, two of them share (to different degrees) code derived from the same fire model used in ELM. (It was not great to learn that only in the Discussion, by the way—this is an important caveat that should have been highlighted before or perhaps in the Results section.)

The four models were obtained from FireMIP phase II (ISIMIP3a) (https://protocol.isimip.org/#/ISIMIP3a/fire). By the time this research was performed, only these four models were available. However, we have now updated our analysis to include a total of seven ISIMIP3a models used in the latest ISIMIP3a benchmarking study (Burton et al. 2024). For instance, the spatial map comparison (Figure 4 in revision) has been updated as follows. The different fire models used by each ISIMIP3a model are now described in the method section 2.2.1.

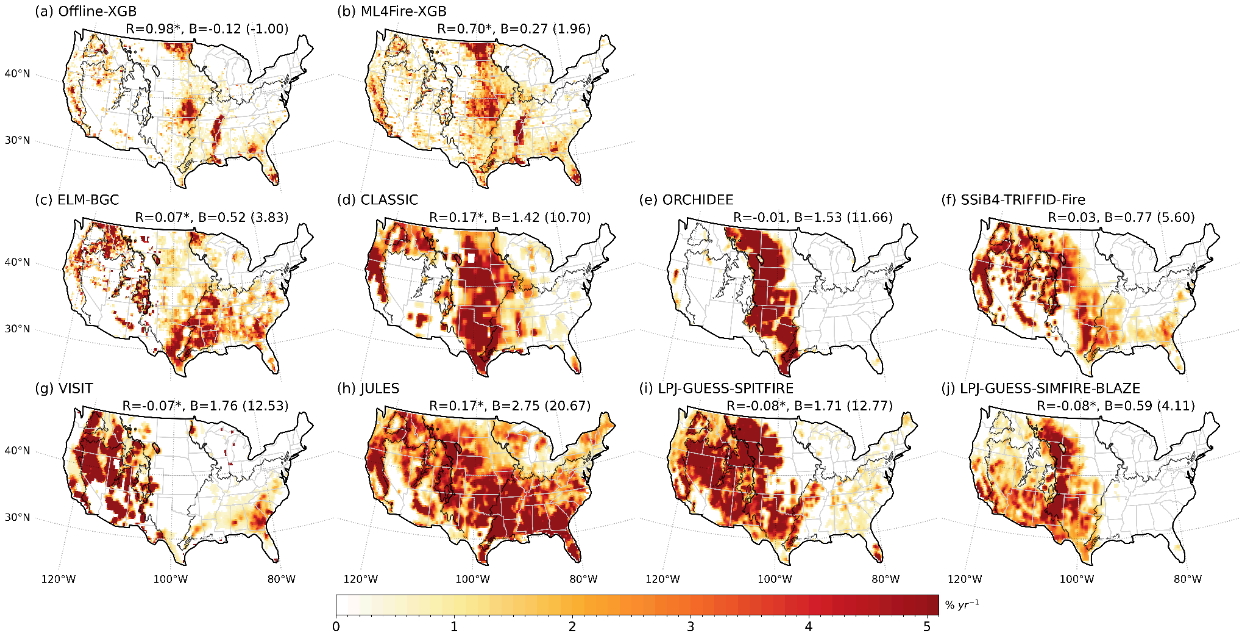


Figure 4. Same as Figure 3, but shows model outputs. The bias (B) against GFED5 is indicated.

Reference:

Burton, C., Lampe, S., Kelley, D. I., Thiery, W., Hantson, S., Christidis, N., Gudmundsson, L., Forrest, M., Burke, E., Chang, J., Huang, H., Ito, A., Kou-Giesbrecht, S., Lasslop, G., Li, W., Nieradzik, L., Li, F., Chen, Y., Randerson, J., Reyer, C. P. O., and Mengel, M.: Global burned area increasingly explained by climate change, Nat Clim Change, 10.1038/s41558-024-02140-w, 2024.

• FireMIP does its own benchmarking at the global scale. If only choosing a few models, their performance in that global benchmarking should be discussed. That would help contextualize your CONUS results.

Thank you for the suggestion. The global benchmarking performance of the models was thoroughly discussed by Burton et al. (2024). To avoid redundancy, we focused our analysis on the CONUS rather than repeating the global evaluation. In the revision, we have included all seven models and discussed their performance over the CONUS.

- P16 L60–61: “VISIT adopts the Thonicke et al. (2001), a semi-empirical fire model and has not been well calibrated since coupling.” According to whom? Or is this just speculation?

This sentence has been written as follow:

VISIT employs the semi-empirical fire model developed by Thonicke et al. (2001), primarily to predict fire emissions. Burned area is calculated annually without accounting for specific ignition mechanisms. This annual burned area is then distributed across each month by weighting it according to the ratio of monthly fire season length to annual fire season length.

Agricultural burning

- It seems like crop vegetation patches and/or burning are included in your model training and analyses? This is worth mentioning, because some process-based fire models exclude crop burning, the detection of which was a major development in GFED5. And crops had by far the largest CONUS (GFED region TENA) burned area in GFED5 (Chen etal., 2023, Table 3), although I’m not sure how much they contributed to the increase from GFED4s. Please discuss.

The crop PFT fraction is included as a predictor in our model, and crop burning is incorporated within the GFED5 burned area data used for training. Consequently, our XGB model is capable of predicting agricultural burning patterns. This inclusion is important because crop burning constitutes 49% of the total burned area in the CONUS, as highlighted by GFED5 (Chen et al., 2023). By accounting for crop burning, our model aligns more closely with recent advancements in fire detection and provides a more comprehensive representation of fire activity across different land cover types, including agricultural areas. We have added the following discussion in Line 363-366.

In contrast, for instance, the ML model includes the crop PFT fraction and accounts for agricultural burning in its training data, enabling it to capture agricultural burning patterns that are typically missing or underrepresented in process-based models. This inclusion is particularly significant in the CONUS, where agricultural burning constitutes 49% of the total burned area (Chen et al., 2023).

- Some of the FireMIP models you chose might have also excluded pasture burning; see Table S3 in Rabin et al. (2017), although that information might not apply to the versions of the models in the FireMIP simulations you chose (see my comment about P6 L43). Indeed, you acknowledge that models might not include cropland or pasture burning at P10 L50–53.

Thank you for raising this point. None of these models explicitly accounts for crop (residual) fires. Most models, except JULES, consider croplands as non-burnable. JULES treats cropland similarly to natural grassland, while all other models exclude cropland from burning entirely. Fires are permitted in pastures across all models. In LPJ-GUESS-SIMFIRE-BLAZE, pastures are harvested, which results in reduced biomass and, consequently, a smaller burned area. In contrast, other models treat pastures as natural grasslands in terms of growth and fire behavior.

For more details, please refer to Extended Data Table 1 (Fire model overview) and Section 3 of the Supplementary Material in Burton et al. (2024), and Teckentrup et al., 2019.

An introduction of the current treatment of crop fire in the current ISIMIP3a models has been added in Lines 159-164.

The representation of fires over croplands and pastures varies across models (Burton et al. 2024, Teckentrup et al. 2019). Most models, except for JULES, classify croplands as non-burnable. JULES treats croplands similarly to natural grasslands regarding fire behavior, while all other models exclude croplands from burning. Fires are allowed in pastures in all models in terms of both growth and fire behavior. In LPJ-GUESS-SIMFIRE-BLAZE, pastures are harvested, leading to reduced biomass and consequently a smaller burned area. In contrast, other models treat pastures as natural grasslands.

Reference:

Teckentrup, L., Harrison, S. P., Hantson, S., Heil, A., Melton, J. R., Forrest, M., Li, F., Yue, C., Arneth, A., Hickler, T., Sitch, S., and Lasslop, G.: Response of simulated burned area to historical changes in environmental and anthropogenic factors: a comparison of seven fire models, Biogeosciences, 16, 3883-3910, 2019.

• P10 L52–53:

Citations should be provided for none of the models having cropland fire on.

References (Burton et al. 2024) and (Teckentrup et al. 2019) have been added. Thank you.

"That says all vegetation models treat pastures as natural grasslands." (a) What is "That" referring to? (b) Citations? (c) For fire only, or are these also not grazed?

We will clarify this in the next manuscript in Lines 276-280.

As noted by Teckentrup et al. (2019) and Burton et al. (2024), none of the process-based models has activated the explicit cropland fire model. Fires are allowed in pastures. While LPJ-GUESS-SIMFIRE-BLAZE incorporates harvesting in pastures, reducing biomass and influencing fire dynamics, all other process-based vegetation models treat pastures as natural grasslands for both vegetation growth and fire processes.

• P10 L53–55: "This may explain the significant overestimation of burned areas in ORCHIDEE as the SPITFIRE fire module has a much higher flammability in natural grasslands compared to woody plants." That suggestion implies that grass isn't in reality more flammable than woody plants, which I don't think is supported by evidence. Perhaps change this to something about grass being TOO much more flammable than woody plants.

Thank you for pointing this out. We agree that in ORCHIDEE, the flammability of grass might be set too high relative to that of trees, as discussed in Teckentrup et al. (2019). We have revised this sentence in the updated manuscript to clarify this point.

• P10 L55–57:

Clarify that “fuel properties” includes amount as well as physical (e.g. bulk density) and chemical characteristics.

Modified as suggested.

Management should also be mentioned here, both in terms of grazing (impacts on fuel load and plant community composition) as well as prescribed fire.

The following sentence has been added to Lines 284-285.

Fuel management practices, such as prescribed burning and grazing, can significantly alter fire dynamics but are generally absent in current models.

• P12 L75–84: The fire model in CLM (which ELM is based on) includes crop fires. Are those not simulated in ELM? Or is their area just low (or even zero) in the Great Plains relative to other types of fire?

Crop fires are not enabled in the version of the ELM model we used. The crop model has not been explicitly calibrated to represent crop fires in CONUS, and enabling it could introduce additional biases due to parameter uncertainties.

“expect” should be “except”.

Corrected.

• P12 L85–88: This sentence should be expanded to explicitly mention and cite process-based models that do have managed crop and/or pasture burning.

To the best of our knowledge, ELM is one of the few process-based models capable of explicitly simulating crop fires; however, this feature was not enabled in our study. None of the models used here include explicit representations of pasture burning. We have added this statement to the revision.

Minor comments

• I would probably remove “North American” from the title, since this work is actually limited to the continental US—well less than half of North America.

“North American” has been removed from the title.

• Title says “ML4Fire-XGBv1.0,” but “ML4Fire-XGB” is used in the paper only to refer to the uncoupled ML-based fire model(s). The real development in the paper is really more about coupling the ML fire models with ELM. I’d suggest changing the model/version number in the title to something like “ELM2.1-XGBfire1.0”.

It’s a great point. We have changed the title to:

ELM2.1-XGBfire1.0: Improving wildfire prediction by integrating a machine-learning fire model in a land surface model

• Line numbers seem to just show the last two digits; please fix in revision.

Looks like it happened when converting from Word to PDF. It has been corrected in the revision. Sorry for the inconvenience.

• Various places: "COUNS" typo.

Apologies for the typos. We have thoroughly gone through the manuscript to avoid these typos.

• P2 L27–29: "[C]limate change has contributed to a 16% increase in the global burned area over the past two decades, while human influences, including ignition and suppression, have reduced by 27%." Second part is sort of ambiguous. Has the strength of the human influences decreased by 27%, or have human influences caused burned area to be reduced by 27%?

R: We apologize for the confusion. We have rewritten this sentence as follow (Lines 27-31).

Globally, modeling studies show that climate change since the early 1900s has contributed to a 16% increase in the total burned area; however, human activities have led to a 19% decrease, resulting in a slight net decline in burned area over the 20th century (Burton et al. 2024). In the past two decades, satellite-derived data suggest that the global total burned area has declined by over 20%, with this trend primarily attributed to human influences (Jones et al. 2022; Andela et al. 2017).

• P3 L67–69: "The corresponding changes in fire dynamics may shift the vegetation species distribution from those originally low in resistance to wildfire to those in high resistance or even benefiting from regular fire occurrence (Rogers et al., 2015; Huang et al., 2024)." Since you're using big leaf, you're not getting that—this should be discussed.

Currently, ELM is configured in the "biogeochemistry" (BGC) model, with PFT distributions prescribed based on satellite products. We have clarified this in the discussion of the revised manuscript in lines 394-396.

• P4 L92–93: Is suppression not also a function of population density (in addition to GDP)? Per-capita GDP, no?

Thank you for pointing this out. Yes. Suppression is parameterized as a function of GDP per capita and population density in ELM. Correction has been made in the revision.

• P4 L04: Worth pointing out that "competition" in ELM (without FATES turned on, that is) is limited to competition for soil resources, not light.

We have clarified it in the revision, Lines 105-106.

The post-fire vegetation recovery in ELM-BGC depends on the plant photosynthesis processes and PFT competition strategy for soil resources.

- P5 L16: “To reduce overfitting, we build a separate ML model for each year from 2001 to 2020 using the remaining 19 years’ data.” Confused me for a while. Suggested revision in bold: “To reduce overfitting, we build a separate ML model for each year from 2001 to 2020 using the data from the other 19 years in that period.”

This sentence has been removed since the random splitting is used to build a canonical model.

- P6 L40–41: “significant” should be “important” or something similar; “while all zero burned areas” seems to be an incomplete thought.

This sentence has been rewritten as follows (Lines 200-201). Thank you.

This step is important to avoid feeding the ML model distinct predictor combinations that all correspond to zero burned areas, which could skew the model's learning process.

- P6 L42–48: Also mention here that you’ll be looking at ELM-BGC outputs.

The ELM-BGC has been added.

- P6 L62: Missing degree symbol at “0.25x0.25”.

Corrected.

- P6 L62: How were the datasets resampled? Nearest-neighbor?

All variables are interpolated using the bilinear interpolation method for spatial and nearest-neighbor for temporal.

- Table 1: GDP and population density citations don’t match text at P6 L61–62.

Corrected.

- P8 L97: “XBG” should be “XGB”.

Corrected. Thank you.

- Fig. 3:

What is gray?

The gray shading has been removed in the revised manuscript.

Would it be more useful to have the ecoregions overlaid on this map instead of state boundaries? I could see an argument either way. Please add text boxes with each model’s Rp (including asterisks to show significance level) and bias scores. As it is now, some models don’t have their scores listed anywhere in the text/figures/tables.

We have updated Figures 3 and 4 in the revision as suggested. Please see the figures in response to your major comments.

- P9 L22–23: “the performance holds,” but it actually worsens, as the next sentence says.

Thanks for pointing this out. This sentence has been changed to:

While integrated with ELM, the performance was slightly degraded.

- P10 L52: “intermodal” should be “intermodal”.

We changed “intermodal” to “inter-model".

- P10–11 L57–59: If fires in this region are managed by prescribed burning, we actually shouldn’t expect the process-based models to do well there, since they don’t account for prescribed burning. This result is thus somewhat surprising.

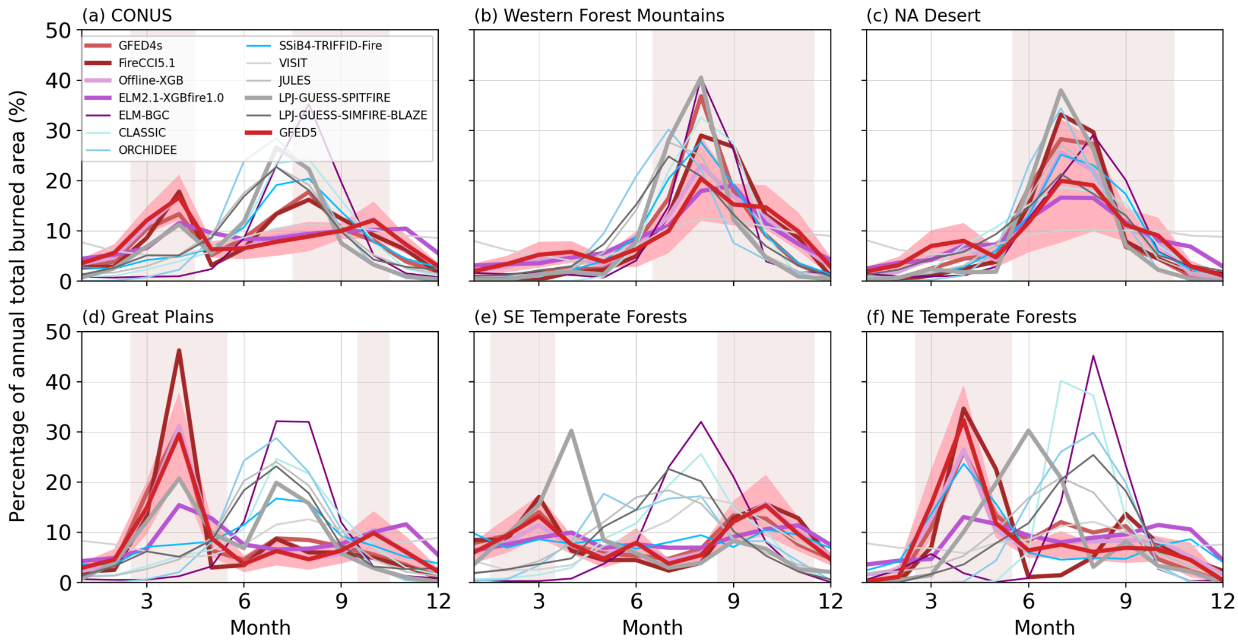
It is a great point. Based on our conversation with local agencies, they tend to cast prescribed fires as much as possible, which effectively reduces large fires and makes the actual burned areas largely influenced by climate and fuel conditions. Since the ignition is less constrained, the burned area is mainly influenced by the fire spread which is highly related to natural forcing such as fuel and wind conditions. On the other hand, the southeastern U.S. is a lightning-prone region, which is a major source of fire ignition in models. Therefore, without prescribed burning, models simulated a high level of fire ignition due to lightning, and well capture the fire spread.

The following discussion has been added in Lines 288-290:

Although prescribed burning as an additional ignition source is not included in the process-based models, ignition is not a limiting factor in this region due to the abundance of lightning, which provides sufficient natural ignition sources. Consequently, burned area is primarily controlled by fire spread, which is influenced by natural conditions such as fuel availability and wind, allowing the models to perform well in simulating fire dynamics.

- P11 Fig. 4: To reflect interannual variability, add uncertainty bars and/or change the red line to a shaded region.

One-standard deviation range of the red line (GFED5) has been added to reflect the uncertainty.



• P13 L93: “accounting for” should be “representing”; “largely” should be “greatly”; “in” should be “from”.

Corrected. Thank you.

• P13 L99: “IVA” should be “IAV”.

Corrected.

• P13 L11–13: “Again, climatic factors play a dominant role in shaping the temporal variability of BAF in the WUS, while human activities largely influence the BAF in the Great Plains and EUS. Process-based models tend to better describe responses of fuel load and combustibility to climate than responses of fire ignition and suppression to human activities.” Citations needed for these statements.

Citations including Kupfer et al. 2020, Chen et al. 2023, and Hantson et al. 2016 have been added.

Kupfer, J. A., Terando, A. J., Gao, P., Teske, C., and Kevin Hiers, J.: Climate change projected to reduce prescribed burning opportunities in the south-eastern United States, Int. J. Wildland Fire, 29, 764–778, 2020.

Chen, Y., Hall, J., van Wees, D., Andela, N., Hantson, S., Giglio, L., van der Werf, G. R., Morton, D. C., and Randerson, J. T.: Global fire emissions database (GFED5) burned area, https://doi.org/10.5281/ZENODO.7668423, 2023.

Hantson, S., Arneth, A., Harrison, S. P., Kelley, D. I., Prentice, I. C., Rabin, S. S., Archibald, S., Mouillot, F., Arnold, S. R., Artaxo, P., Bachelet, D., Ciais, P., Forrest, M., Friedlingstein, P., Hickler, T., Kaplan, J. O., Kloster, S., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Melton, J. R., Meyn, A., Sitch, S., Spessa, A., van der Werf, G. R., Voulgarakis, A., and Yue, C.: The status and challenge of global fire modelling, Biogeosciences, 13, 3359–3375, 2016.

- P15, Fig. 8: What is white?

This figure has been removed.

- Two different reference lists? The first one ends and the second begins on P19. They're not the same, either, with at least one reference (Donovan et al., 2020) missing from the first.

Thanks for checking that! We have corrected the reference list in the revision.