Review: ML4Fire-XGBv1.0: Improving North American wildfire prediction by integrating a machine-learning fire model in a land surface model (gmd-2024-151)

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***.

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).

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.

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).

    - 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?
    - 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.
- 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.
- 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.

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.
- 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).
- 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!

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.)
- 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.
- 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?

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 et al., 2023, Table 3), although I'm not sure how much they contributed to the increase from GFED4s. Please discuss.
- 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.
- P10 L52–53:
    - Citations should be provided for none of the models having cropland fire on.
    - "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?
- 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.
- P10 L55–57:
    - Clarify that "fuel properties" includes amount as well as physical (e.g. bulk density) and chemical characteristics.
    - Management should also be mentioned here, both in terms of grazing (impacts on fuel load and plant community composition) as well as prescribed fire.
- 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?
    - "expect" should be "except".
- 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.

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.
- 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".

- Line numbers seem to just show the last two digits; please fix in revision.
- Various places: “COUNS” typo.
- 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%?
- 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.
- P4 L92–93:
    - Is suppression not also a function of population density (in addition to GDP)?
    - Per-capita GDP, no?
- P4 L04: Worth pointing out that “competition” in ELM (without FATES turned on, that is) is limited to competition for soil resources, not light.
- 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**.”
- P6 L40–41: “significant” should be “important” or something similar; “while all zero burned areas” seems to be an incomplete thought.
- P6 L42–48: Also mention here that you’ll be looking at ELM-BGC outputs.
- P6 L62: Missing degree symbol at “0.25x0.25”.
- P6 L62: How were the datasets resampled? Nearest-neighbor?
- Table 1: GDP and population density citations don’t match text at P6 L61–62.
- P8 L97: “XBG” should be “XGB”.
- Fig. 3:
    - What is gray?
    - 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 R_p (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.
- P9 L22–23: “the performance holds,” but it actually worsens, as the next sentence says.
- P10 L52: “intermodal” should be “intermodal”.
- 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.
- P11 Fig. 4: To reflect interannual variability, add uncertainty bars and/or change the red line to a shaded region.
- P13 L93: “accounting for” should be “representing”; “largely” should be “greatly”; “in” should be “from”.

- P13 L99: “IVA” should be “IAV”.
- 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.
- P15, Fig. 8: What is white?
- 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.