(Note: The reviewer's comments are in gray and the author's responses are in blue. Unless specified otherwise, the line numbers quoted in our responses are with reference to the revised manuscript.)

Overall Feedback

This is another work applying ML to predict fires, but this time more focusing on the WUS where fires are more connected to human activities. This take on the ML-based fire model is unique and interesting. The ML framework used here is standard, though the advantage of SMLFire1.0 compared to the gradient-boosted tree model needs to be more explicitly described and discussed in the main text. For instance, one of the advantages of SMLFire1.0 is uncertainty quantification. What source of uncertainty is evaluated? If it is only statistical uncertainty, would it be model-dependent? How does this model account for the spatiotemporal variability of the predictors and their non-linear interactions and why the other models couldn't? The cited work Wang et al., 2021 also considered those features in their XGBoost model and SHAP analysis. The interpolation of SHAP values also needs to be more carefully reviewed. Such as human predictors can introduce contrast impacts even at the same location. Higher population density can cause both fire ignition and suppression, a small SHAP value would overlook the impact of this predictor. For these reasons, and others mentioned below, major revisions to the manuscript are needed before possible publication.

We thank the reviewer for carefully reading through our manuscript and providing constructive feedback on our ML framework as well as its interpretation through SHAP values.

One major point raised by the reviewer is regarding the comparison of SMLFire1.0 with the gradient-boosted tree model of Wang et. al., 2021. While the Wang et. al., 2021 model also accounts for the spatiotemporal variability of predictors and their nonlinear interactions, we identify two major differences, namely their gradient-boosted tree model: a) only predicts the area burned at the grid cell level and not the fire probability or frequency, thus missing an important element of stochasticity in fire activity; b) does not perform uncertainty quantification, while our model uses the variance of Monte Carlo samples from the optimized mixture model to estimate the parametric model uncertainty for fire frequency and sizes. Although we view the analysis of Wang et. al., 2021 as complementary to ours, we make the distinction between our approaches sharper in the Theory section as follows:

(lines 184-191)

> *c) be based on parametric distributions that could be sampled using Monte Carlo simulations for estimating the mean and parametric model uncertainty of modeled fire frequency and sizes. While tree-based ML approaches using xGBoost have shown high performance in area burned prediction across the continental US (Wang et al., 2021), we adopt a neural network based architecture here because it combines the flexibility of machine learning techniques with the robustness of parametric distribution based methods traditionally used in statistical fire modeling (Westerling et al., 2011, Joseph et al., 2019). Moreover, since neural network models have more powerful representation*

learning capabilities than gradient-boosted trees (Levin et al., 2022), they are better equipped for generalizing the learned relationships between input predictors and fires to test data from future climate states or different fire regimes.

(lines 227-228):

We treat the variance as an estimate of the parametric model uncertainty, or equivalently the uncertainty in modeled frequency due to different realizations of a parametric model.

The first paragraph of the Conclusions section (lines 608-616) has also been lightly edited to make the point about parametric model uncertainty estimation clearer.

We address the reviewer's comments regarding the interpretation of SHAP values and possible confounders in the section on "Explanation of SHAP values" below.

Major Remarks

Selection of predictors

1. The meteorological predictors were obtained from multiple data sources. The inconsistency between the data sources may introduce additional uncertainties. Why are the daily and X-day minimum and maximum temperatures extracted from different sources? The fire weather index is calculated from relative humidity and wind speed from gridMET. Both variables from UCLA-ERA5 were used as individual predictors. Why not use the same data source to derive FFWI?

We thank the reviewer for raising these points. Firstly, we used the NOAA nClimgrid data to obtain the monthly means of daily maximum and minimum temperature data, which is only available on monthly (and not daily) scales. We preferred the NOAA data over gridMET because it is of higher quality and extends back to 1895, which will enable us to validate our model on fire data from the early 20th century in forthcoming work.

In order to calculate the X-day daily maximum and minimum temperatures we used daily scale data from the UCLA-ERA5 reanalysis. More importantly, we used temperature, humidity, and wind speed data from UCLA-ERA5 reanalysis to calculate the monthly mean FFWI and X-day maximum FFWI. We acknowledge an unfortunate error in the original manuscript (line 124) where we stated that FFWI data was taken from gridMET. The only predictor obtained from gridMET in our analysis is the monthly mean FM1000 value.

We could have used daily scale data from gridMET for all the above predictors besides FM1000, however since wind speed in gridMET is derived by downscaling NARR data from a coarser resolution of 32 km x 32 km resolution, we instead used wind speeds from UCLA-ERA5 reanalysis which downscaled ERA5 wind data to a higher 9 km x 9 km resolution using the WRF model. Figure 11 in Rahimi et. al., 2022 illustrates the improvement in wind speed resolution due to UCLA ERA5-WRF as compared to gridMET-NARR. In order to ensure minimal error from using multiple data sources as the reviewer has pointed out, all X-day maximum predictors as well as monthly mean FFWI have been derived using data from UCLA ERA5-WRF reanalysis.

Following is the revised data description in the manuscript (lines 123-129):

> *Monthly mean FM1000 values, an indicator of climate-derived moisture balance, were adapted from gridMET (Abatzoglou, 2013).The FFWI, which is calculated using temperature, humidity, and wind speed (Fosberg, 1978), has been shown to be an important correlate of dry, windy conditions associated with fire weather (Moritz et al., 2010). Since wind speed in gridMET is derived using a spatial interpolation of the National Atmospheric Regional Reanalysis (NARR) data from a coarser (32 km x 32 km) resolution, we instead use high (9 km x 9 km) resolution temperature, humidity, and wind speed predictors from the dynamically downscaled UCLA ERA5-WRF reanalysis (Rahimi et al., 2022) to calculate the monthly mean FFWI.*

2. In addition to monthly mean daily maximum and minimum temperature, what physical information can the X-day mean variables add in? Is X-day mean calculated from the running average?

We have expanded the discussion around the current description of X-day mean variables and also commented upon the physical meaning behind their inclusion (lines 129-132):

> *Furthermore, we use daily scale data from the UCLA ERA5-WRF reanalysis to calculate the monthly maximum X-day running average of daily maximum and minimum temperature ($Tmax^{maxX}$, $Tmin^{maxX}$), where $X \in \{3, 5, 7\}$. Similar X-day extreme predictors are also derived for VPD, FFWI, and wind speed. The X-day running average of these predictors are included to improve our model's sensitivity to weekly scale extreme fire weather caused by events such as heatwaves.*

3. The predictors are selected with physical meanings. Could the authors elaborate on why a variable is chosen?

Besides adding an explanation for including FM1000, FFWI, and X-day extreme weather variables in the Data section, we have now included an additional table in the Supplementary information, Table S3, that elaborates briefly on the relationship between each predictor and fire response variables.

4. Line 162: Table S2 lists 30 predictors. This number of predictors is more close to the one after iteratively dropping off predictors that do not improve overall performance and are highly correlated as the authors mentioned in the results section. If two variables are highly correlated, which one will be kept? I would also move this to the method section.

We apologize for the confusion regarding the number of variables in Table S2. We have added an extra line in the caption explaining how the total number of potential predictors in Table S2 adds up to 51:

Considering each predictor's M antecedent months' average and maximum X-day running average components as distinct predictors, the total number of predictors adds up to 51.

As per the reviewer's suggestion, we have also moved the discussion regarding variable selection to the Methods section (lines 316-320).

5. Since the authors found that the antecedent precipitation is one of the most important drivers affecting plant growth, would it be helpful to include vegetation predictors that more closely connect to fuel conditions? Meanwhile, the results show low importance of the spatial variability of vegetation predictors, however, the temporal variability is important, which appears to conflict with the justification of using time-invariant biomass.

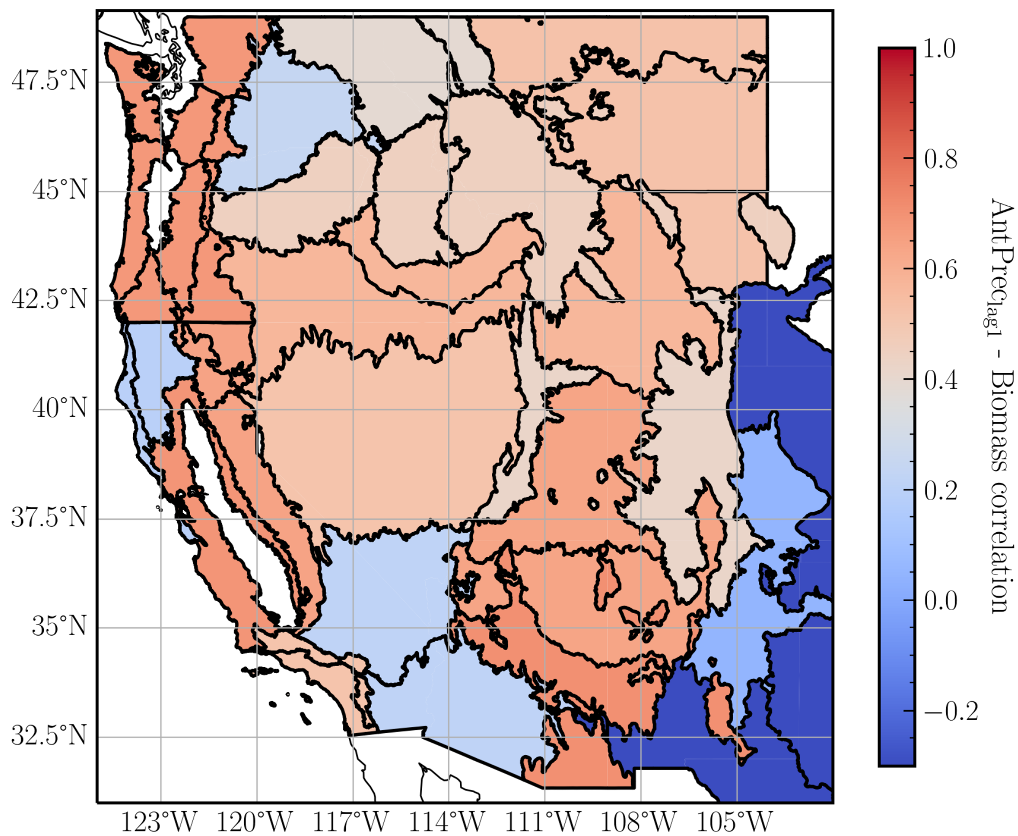


Figure 1. Spatial map of the correlation between the mean of antecedent precipitation in the previous year and aboveground biomass aggregated for each Level III ecoregion. The boundaries of various ecoregions are indicated by solid black lines.

The reviewer raises an important point regarding the importance of temporal variability of vegetation predictors. We acknowledge that the abundance and flammability of fine fuels is one of the main drivers of fire activity. Time-varying biomass maps would be an ideal predictor for our model; however, to the best of our knowledge, no such products exist in the fire ecology or remote sensing literature.

Thus, we use a combination of both: a static biomass map to inform the model of relative fuel abundance as well as antecedent precipitation predictions in lag years as a proxy of dynamic plant growth. Interestingly, as we show in Fig. 1 above, the long term mean of $AntPrec_{lag1}$ shows moderate correlations with the time-invariant aboveground biomass across different L3 Ecoregions.

We also note from Fig. S2, S3 and S7, S8 in our Supplementary Information that $AntPrec_{lag1}$ is an important predictor for fire frequency while Biomass is an important predictor for fire size. A potential explanation could be that while $AntPrec_{lag1}$ is responsible for plant growth that drives fire frequency in arid climates, the spatial fuel abundance given by Biomass, especially in Forests, is a major driver of fire spread resulting in larger burned areas.

6. Surprisingly, human predictors are not among the top 10 predictors for fire frequency, while over 90% of the California fire ignitions were associated with human activities (e.g. Balch et al., 2017).

This is an insightful comment and is related to point 3 of Referee Comment (RC) #3 on our original preprint. Before outlining how we addressed the reviewer's comment, we note that the FPA-FOD data used in Balch et. al., 2017 contains significantly more smaller fires than the WUMI dataset that we are using which only contains fire ≥ 1 km^2. Since human fires are mostly small fires, we expect the ~90% number for human started fires in California to be slightly lower for the WUMI dataset.

There are two potential explanations for the lack of human predictors among top 10 predictors for fire frequency: a) since our model is trained on fires across the WUS where the proportion of human started fires is ≤ 50%, there could be a potential skewed sampling of fires while training SMLFire1.0 (note that we do not have access to labels for human vs natural ignitions in our data set); b) our choices of human predictors are poor correlated with fire occurrences.

As outlined in RC#3, we also performed a toy experiment to test the explanations outlined above. Based on our findings, we have added the following sentences to our discussion clarifying the potential bias in SMLFire while modeling fires with human vs natural ignitions:

(lines 431-434) *Given that a large fraction of fires in parts of the WUS, especially Mediterranean California and coastal PNW (Balch et. al. 2017), are human ignited, this result could stem from a skewed sampling of fires while training SMLFire1.0 as well as the lack of correlation between our chosen human predictors and fire occurrences.*

(lines 568-570) *Alternatively, we could leverage the seasonal differences between human and lightning started fires to account for potential selection biases in training data for SMLFire1.0.*

Explanation of SHAP values

1. Line 303-304 As far as I know, the Kernel SHAP also makes an assumption about feature independence. If two features are highly correlated, one value will be replaced with random

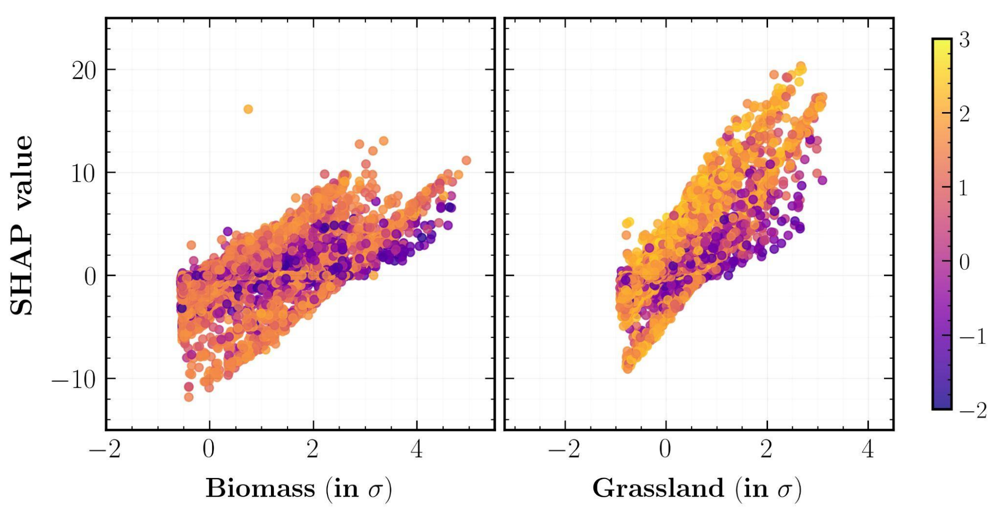
ones from the background dataset, and then SHAP will generate predictions based on the new datasets while making the SHAP value estimation less reliable.

We agree with the reviewer's comment that the Kernel SHAP method assumes feature independence while calculating the Shapley additive value of a particular feature and also relies on random draws from the dataset for replacing absent features. Thus, in case there are correlated features, Kernel SHAP could end up overweighting the importance of unlikely data points. We have modified our writing in the 'Predictor Importance' subsection to clarify this point (lines 325-329):

> *This is in contrast to the traditional predictor importance techniques which only rely on a fixed coalition of predictors to assess the contribution of an individual variable … We note that a drawback of using the Kernel Explainer method is that its assumption of predictor independence could lead, in practice, to a biased estimation of predictor importance in the presence of two or more strongly correlated features.*

2. Figure 11-Forests shows increases in both grassland fraction and above-ground biomass increase the burned area. In the forests, the above-ground biomass is mainly contributed by wooden biomass. Therefore, I am wondering if grassland fraction and biomass will increase simultaneously. Could the authors plot a partial dependence plot for those two predictors? How to understand the higher Tmax would suppress fire spread? When replacing VPD with relative humidity relevant variables, the relationship between Tmax and burned area becomes positive (Figure S10-lower left panel). Would the correlation between the predictors affect the results?

We find that the Ecoregions are the more appropriate spatial scale to assess the relative predictor importance of Biomass and Grassland. Analyzing Figs. S7 and S8, it is clear that the cumulative importance of Grassland at the Forests Division level is primarily driven by its role in CA Central and South Coasts, Southern Rockies, and AZ/NM Mountains Ecoregions, whereas Biomass is the more important predictor in Sierra Nevada, CA North Coast, PNW Mountains, and Northern Rockies. For completeness, we also include a plot comparing the partial dependence plot for Biomass and Grassland for the aggregated Forests Division:



Indeed, as the reviewer's previous comment suggests, correlation between VPD and Tmax appears to affect the predictor importance of Tmax in Forests (Fig. 13) relative to the case with only RH and Tmax (Fig. S10; lower left panel). We have now added an additional line in our Results section clarifying this point (lines 542-545):

> *From the perspective of predictor importance, there might actually be an advantage to using RH instead of VPD: the correlation between VPD and Tmax leads to a small but spurious trend in SHAP value for Tmax in Forests as shown in Fig. 13, whereas using only RH and Tmax in Fig. S10 yields the correct Tmax effect on fire size. On the other hand, from the perspective of future climate-fire relationships, …*

Specific Remarks

1. A stochastic ML fire model is introduced. I am wondering if any physical processes (e.g., lighting and human ignition) are included in this model?

We have included lightning the lightning strike density as a correlate of natural ignitions and several human predictors for human ignitions. However, unlike process-based models such as SPITFIRE (Thonicke et. al., 2010), we do not include any physical parameterizations for ignitions. We appreciate the reviewer's suggestion here and will consider including additional physics in SMLFire1.0 to improve its ignition modeling capacity.

We have also included two additional sentences in the Data section to clarify the inclusion of human predictors and the dual role played by predictors such as population density (lines 166-168):

> *These predictors serve as potential correlates of human ignitions for fire occurrences as well as proxies for access to fire suppression or containment resources. Some predictors such as Popdensity could play a dual role through both increasing the likelihood of ignitions while also providing easier access for fire suppression.*

2. Mapping the Ecoregions also on Figure 1 would much help the reader to understand their locations. The "desert" is not a place closely related to fires, considering using an alternative name for this division.

We acknowledge the scope for confusion due to our use of "desert" to describe the arid grassland and shrubland regions spanning several Ecoregions through the center of our study region. However, we chose "desert" following the nomenclature used by the EPA while referring to the Level II ecoregion, North American Cold and Warm Deserts, containing the area roughly equivalent to our Deserts Division.

3. Line 125: What's the spatial resolution of ERA5-WRF? Suggest adding spatial resolution to Table S2.

We have added spatial resolution to Table S2 as an additional column, and also included a clarification in the caption about how all predictors, despite having different native resolutions, are aggregated to the 12 km resolution in our statistical analysis.

4. Line 153: The variable Popdensity actually measures the distance to human settlements. However, the Popdensity sounds like population density which is inversely proportional to the distance. Suggest changing to Pop_dist or similar terms.

We have modified all references to 'Popdensity' throughout the text as well as within plots to 'Pop10_dist' following the reviewer's suggestion.

5. Line 221-224: While I am okay with this approximating, the MTBS dataset provides the extent of fires 1000 acres or greater in the WUS, which might be more precise than approximating each fire as a circle.

We agree with the reviewer's comment; the main consideration for approximating the fire shape as a circle was driven by the absence of burned area polygons in the WUMI dataset, which contains both interagency and MTBS fires. We have added the following sentence (lines 241-242) in the Methods section to clarify the reviewer's comment:

> *In future work, we will use burned area polygons from MTBS for large fires instead of the circular approximation while deriving the effective input predictors.*

6. Line 235-238: Can the MDN predict burned area at each grid cell directory or for individual fires? If yes, why not sum up burned area at each grid cell or individual fires to get the Ecoregion level burned area?

We derive the Ecoregion level burned area by summing up the areas for individual fires – exactly as the reviewer has suggested (see lines 249-250 in the revised manuscript). In the referenced passage, we have outlined a simple analytic expression for estimating the relative contribution from frequency as well as fire sizes to the mean burned area at a given scale. For instance, if we were to use a ML model for fire frequency and overpredict the number of fires but underpredict individual fire sizes, we would still get the correct burned area. Using the analytic expression provides a helpful diagnostic tool in such a scenario.

We have lightly edited our writing around this point in the revised manuscript to emphasize that the burned area is calculated by summing up individual fires, and the average burned area calculation is only for interpreting our calculation schematically.

7. Line 235: What does the spatialtemporal scale mean here? How can the CCDF plot support the breakpoint selection based on this definition?

As we described above, the use of spatiotemporal scale here is merely schematic. The breakpoint selection procedure is based on creating the CCDF plot with individual fire sizes and not cumulative burned area.

8. Line 242: Should this be “consecutive time period”? Please describe how a breakpoint will be determined in the method.

We have outlined our procedure for determining the breakpoint year as well as different validation steps in the Results section (lines 456 - 465). A similar version of the text can be found in the preprint (lines 435-444) but may have been obscured due to the unfortunate page break due to the placement of Figs. 9 and 10.

9. Line 310-315: As stated in Line 189, 17489 grid cells correspond to active fires. Is it associated with the same value as stated here “˜20000” test points?

To construct the set of test points for our SHAP values, we consider all (i.e 17,489) grid cells with a fire and combine them in a ratio of 1:3 with a random sample of background points with no fires (i.e 17,489/3 ~ 5830), bringing the total number of test points to 23,319 points, or ~20,000 as we mention in the text.

10. Line 361: “are modeled”.

We have fixed this typo in the revised manuscript.

11. Line 451: Which fire frequency is used in the follow on analysis?

Both modeled and observed fire frequency are used to calculate the monthly and annual area burned across the WUS. Another source of stochasticity that we explore in Fig. S6 is deriving the burned area using input predictors corresponding to observed as well as modeled fire locations.

12. Line 449-500: Was this statement based on all the 28 predictors or the top 10 shown in Figure 12? Increasing the distance to the camp ground does not seem to increase the burned area.

We are slightly confused by the reviewer’s comments here since we do not discuss the effect of distance to camp grounds in Fig. 12. The one human related variable that shows up as an important predictor in Fig. 12 is the distance to areas with population density greater than 10 people per square kilometer (Popdensity; or Pop10_dist in the revised manuscript), which indicates that more remote areas are more conducive to larger fire sizes.

13. Line 591-592: The slope variables are also invariant with time.

This is a good point. We have amended lines 591-592 (now lines 613-615) as follows:

> *…our model: relies only on the spatiotemporal variability of dynamic predictors, the spatial variability of static predictors, and not on any predictors related to the location and time such as latitude or calendar month;...*

14. Line 603-604: What does the “fire month" stand for?

We have adopted “fire month” as a descriptor throughout the manuscript to refer to monthly scale predictors for any month with a fire. We mainly use it to distinguish the effect of climate and fire weather conditions during a month from the influence of antecedent climate predictors. We have now clarified this usage in the Data section by modifying the following sentence (line 130) where we first use the term fire month:

> *Thus, for a given month m with potential fire activity (henceforth fire month) …*

References:

1. Balch, J. K., Bradley, B. A., Abatzoglou, J. T., Nagy, R. C., Fusco, E. J., & Mahood, A. L. (2017). Human-started wildfires expand the fire niche across the United States. Proceedings of the National Academy of Sciences, 114(11), 2946–2951. https://doi.org/10.1073/pnas.1617394114

2. Moritz, M. A., Moody, T. J., Krawchuk, M. A., Hughes, M., and Hall, A. (2010), Spatial variation in extreme winds predicts large wildfire locations in chaparral ecosystems, Geophys. Res. Lett., 37, L04801, doi:10.1029/2009GL041735

3. Rahimi, S., Krantz, W., Lin, Y.-H., Bass, B., Goldenson, N., Hall, A., et al. (2022). Evaluation of a reanalysis-driven configuration of WRF4 over the western United States from 1980 to 2020. *Journal of Geophysical Research: Atmospheres*, 127, e2021JD035699. https://doi.org/10.1029/2021JD035699

4. Thonicke, K., Spessa, A., Prentice, I. C., Harrison, S. P., Dong, L., & Carmona-Moreno, C. (2010). The influence of vegetation, fire spread and fire behaviour on biomass burning and trace gas emissions: Results from a process-based model. *Biogeosciences*, *7*(6), 1991–2011. https://doi.org/10.5194/bg-7-1991-2010