Response to RC2

We thank the reviewer for taking the time to review our manuscript and for their positive comments.

In the following point-by-point response we have kept the reviewer's comment verbatim in black text, our responses are in blue and proposed new text in green. Line numbers refer to the reviewed version of the manuscript. Text insertions into existing text are additionally denoted with underlining.

The authors present an important contribution to a pertinent issue in wildfire modelling, adding an important perspective to the hot topic of the differences between natural vegetation versus cropland wildfire. The model is well-described and rigorous, finding concrete differences between the drivers of burning in cropland and non-cropland vegetation. Performance statistics are largely better than fireMIP models, but similar-to-worse than the other GLM based studies cited in the introduction. However, as this is the only study focussed on Europe and is at a finer resolution than most of the other studies, the only conclusion that can be reached through this comparison is that BASE's performance meets an acceptable threshold for publication.

Whilst the paper's conclusion emphasises predictive applicability to the projection of future burned area, the overall performance is not greater than existing global and regional modelling methods. The crucial contribution of this paper is the use of a twin-approach to cropland and "natural" vegetation wildfires. The authors demonstrate that this can substantially reduce confounding effects in specific drivers (e.g. population density or fire weather) that differ between land-cover types, and that trends in NCV/cropland fires can be disaggregated. The paper thus reflects an important contribution to the next generation of fire models, towards improving overall model performance and also to understanding the complex dual effects of climate warming and land-use change over time.

Yes, we absolutely agree that whilst we judge the performance of BASE to be sufficient for application, the results about cropland vs NCV fires are the bigger contribution.

Specific Comments:

L95: provide evidence/citation that existing global fire models do in fact perform worse in other biomes due to this training bias. Is this the case for all widely-adopted fire-enable DGVMs?

Good point. We are not aware of any paper that quantifies this (though it is a good idea to write one). Qualitatively, Figure 4 and Table 2. of Hanston et al. 2020 illustrates the nature of the regional discrepancies well. Excluding MC2 and GlobFIRM (which are poor everywhere), all models do a reasonable job in the dry tropics, especially Africa. But model performance is poor and very variable in other regions. All models largely overpredict in extra tropics, most notably in the US (except for ORCHIDEE-SPITFIRE but that models predict very little fire outside of the dry tropics) and the Mediterranean, Middle East and the very arid zones in central Eurasia. Some features of the Eurasian steppe are reproduced by some models, but patterns in the Boreal zone are not well reproduced.

We therefore to propose mention the over prediction in the extra tropics and to support our comment by citing the figure and the table from Hantson et al 2020 in the sentence at line 90:

"However, they have notable regional discrepancies and in particular over-predict burnt area in the extra tropics (see Fig 4 and Table 2 in Hantson et al., 2020), likely because their global focus leaves them unable to resolve regionally-specific processes or phenomena."

L180/L190: explain why MEPI and PHI are defined as relative to GPP maximum month of prior year; as maxima are less stable against interannual variability than, for example, the study-period mean, what was the advantage to this formulation?

Using the maximum to normalise MEPI and PHI is fundamental to the way the indices work. It implicitly accounts for both different seasonal cycles (specifically the length of the growing season) and overall productivity in grid cells. To account for overall productivity, one could use the mean, but this will give numbers higher than 1 which are not so useful in this context and depends on growing season length. The maximum has the advantage that the values will always be between 0 and 1 regardless of growing season length or overall productivity.

To make this a bit more concrete, consider MEPI for an area with a short growing season. If you normalise by the mean, a month with high (near maximum) productivity might have a value of 3. Now consider a month of high productivity in a grid cell with a long growing season, the value might only reach 1.5. But in terms of what them to capture, this difference is not important and actually counterproductive. They are both high productivity months – i.e. not very flammable. So, one grid cell having value of 3 and the other 1.5 introduces a spurious difference between the grid cells and the model will try to fit to that. In fact, we want the same value for all grid cells when they have high (relative) productivity – and that is what using the maximum allows. So, it is essential to use the maximum.

Regarding interannual variability, we did test using a longer time period (25 months instead of 13) to derive the maximum but this didn't improve our results.

So, to give a short explanation to clarify this, we propose to add the following at line 185.

"Using the 13-month maximum accounts for the overall productivity of a grid cell in a manner which is insensitive to the length of the growing season (unlike the annual mean)."

L200/Table 1: In table 1 it is stated that FAPAR12 models fine fuel build up over twelve months. At lower values of fAPAR there is a roughly linear relationship with LAI (LAI ~ - 0.5 ln(1 - fAPAR)). This is only one component of leaf litter accumulation. But this neglects leaf mass per area and leaf lifespan, which can vary substantially between needleleaf/broadleaf/deciduous/evergreen. So what does fAPAR12 actually mean, and why is it selected over the more physical GPP? Could it be that the modelled GPP product does not give as spatially reliable a map as the remotely-sensed fAPAR? This could explain the decreased spatial (but improved interannual) performance when GPP replaces fAPAR (table 2). This issue can either be addressed with a good explanation of FAPAR12's physical effect vs GPP12 in the method section, or by discussing the implications further in section 4.1.3.

This is a good point. The logic behind FAPAR12 was that it should provide a better handle on fine fuel (i.e. leaves) than GPP (some of which will be put into wood), even with the difference in leaf longevity and LMA. But we also tried GPP12 as an alternative, because as you state, it is a little more physical. But then we didn't give too much thought as to why FAPAR12 was the better spatial indicator and gave higher deviance explained even though GPP12 gave better temporal IAV.

Thinking about it now, we think the better spatial performance is probably is indeed because FAPAR12 gives a more direct measure of leaf production (imperfect as it may be). We do have good confidence in the GOSIF GPP product because, being based on SIF and not greenness (like, say MODIS GPP) is much closer to a "direct observation" of GPP. So, we don't wish to attribute the relatively poorer performance of GPP to poorer quality data. We propose to clarify FAPAR12's particular sensitivity to leaf/fine fuel at line 199:

"The fraction of absorbed photosynthetically active radiation (FAPAR) <u>is a proxy for live leaf biomass and</u> can be used to quantify <u>fine</u> fuel buildup and availability (Forkel et al., 2017; Knorr et al., 2016; Kuhn-Régnier et al., 2021)."

And insert at line 475 in the Discussions:

"The better performance of FAPAR12 than GPP12 can be explained by FAPAR's specific relationship to leaf biomass rather than GPP's relationship to general biomass production, and the importance of fine fuel (i.e. leaves) for enabling fire ignition and spread."

Unfortunately, by their construction, the FAPAR12 and GPP12 quantities have both spatial and interannual effects. This makes it very difficult to suggest why GPP12 is better temporally better and prefer not to speculate in this manner. However, it does highlight that there is definitely research to be done here, and that future studies may disentangle this by more carefully considering "spatial only" and "temporal only" predictors.

L214/223: Two sources of GDP data are cited, which was used in this study?

Apologies, this was simply an error in the manuscript. We only used the Kummu data for GDP. We have corrected this.

L478: Could it be that these regions are too wet to seasonally burn?

Yes, indeed that is a reasonable explanation. Ideally the model would account for this directly through the monthly FWI, but this might not fully account for the effect. We propose to add (at line 278):

"or do not have an appropriate burn window due to insufficient precipitation seasonality."

L568: Spain does differ, but there also appears to be more simulated cropland fire in France, Poland and the Baltics. It would be good to either acknowledge this difference, or to justify that this visual difference in the maps is not as significant as the Spanish case (e.g. due to GLM 'smearing' or colormap choice).

Yes, there are overestimates of cropland burning in France, Poland and the Baltics, but these are less than in Spain, and, as the reviewer points out, these are a consequence of the smear of low values from the GLM which are somewhat overemphasised by the threshold in the colour scale. Also, the NCV change is particularly remarkable, we should empathise this more. We propose to clarify by modifying the text at line 559 to read:

"This is particularly clear in the observations when comparing as the observed NCV fires in Spain to neighbouring Portugal (Figure 4). However, BASE NCV fails to simulate this change in fire occurrence at the national border. Furthermore, BASE Cropland also overestimates in Spain, predicting extensive area of cropland burning when in fact there are only limited areas. This overestimate is larger than the low levels of overprediction seen in, for example France and Poland, which is a consequence of the GLM tendency to predict a lot of low values (which is here overemphasized by the threshold in the colour scale). These substantial overestimates of both cropland and NCV fire occurrence may indicate phenomena specific to Spain which are not accounted for in BASE."

L687: Consider changing "so is suitable for projecting changes in fire hazard over annual-to-decadal time scales" to "so is suitable for projecting differing changes in fire hazard between cropland and non-cropland vegetation over annual-to-decadal time scales". Not to say it cannot be used in general projections, but that this is the unique value of the model compared to similarly performing fire models.

Yes, although because we do think the model is generally suitable for projections (especially given its regional focus and higher spatial resolution compared to the global models), we suggest the formulation:

"... so is suitable for projecting changes in fire hazard over annual-to-decadal time scales, particularly when considering cropland and non-cropland land cover types."

Technical Corrections:

L245-250: give the list of actual land cover classes in the supplementary for better readability.

Yes, good idea, we are happy to do this.

L393: add full-stop.

Done.

L490/491: correct "1)" and "2)" to (1) and (2) in sentence.

Done.