Dear reviewers,

Thank you for these comments. They are on the point and highlight useful ways in which the study and manuscript could be improved. We intend to take on board the great majority. In particular, we will:

- *Run scenarios with two additional ESMs and conduct additional factorial experiments to better isolate different drivers;*
- *Add additional statistical rigour to analysis of model outputs and more clearly describe these methods;*
- *Improve the figure resolution!*

We respond to your comments in full below. Specific proposed new text is highlighted yellow.

Reviewer 1

This paper compares future burned area projections from two global fire models — WHAM-INFERNO (focused on socio-economic drivers) and the Haas GLM (focused on ecological drivers) — under SSP1-2.6 and SSP3-7.0, finding that the two models disagree radically on future burned area, particularly regarding the role of direct human forcing. The topic is timely and the comparison is a useful contribution. However, the generalizability is limited by reliance on a single ESM and only two SSPs, and the central claim lacks formal statistical support, together with some other minor issues. I recommend **moderate revision**.

Major Comments

1. Limited scenario and ESM diversity undermines generalizability. The study uses only two SSPs (SSP1-2.6 and SSP3-7.0) forced by a single Earth System Model (UKESM). The authors themselves acknowledge that UKESM projects unusually rapid change in the Amazon (lines 503–505). Using only one ESM makes it impossible to disentangle structural uncertainty in the fire models from forcing uncertainty. At minimum, a middle-of-the-road scenario (SSP2-4.5) and one additional ESM should be included, or the limitation needs much more detailed discussion with clear justification for why this was not done.

Thank you for this comment. This study began as a means to facilitate discussion across fire modelling groups about how different processes are understood and conceptualised, and to develop a combined view of ways ahead. However, we acknowledge that the central finding to emerge from this - of large structural uncertainty in model projections - requires additional substantiation for wider generalizability.

We will add the following runs:

- *We will add 2 additional ESMs from the ISIMIP ensemble (so 3 ESMs for each scenario run). We will use GFDL (lowest equilibrium climate sensitivity of ESMs in ISIMIP) and MRI (median ECS), to complement UKESM (highest ECS).*
- *We will also do additional factorial runs with climate variables held constant (hereafter climate 2020).*
- *We will do specific runs testing the sensitivity of land cover and GPP in the Haas model:*
    - *For land cover, this responds to reviewer comment (2) below.*
    - *For GPP, we will run the climate 2020 experiment with and without GPP held constant. This is because the Haas model conceptualises GPP as its central driving variable, but GPP does not directly impact flammability in WHAM-INFERNO. The models also include many similar biophysical variables (vapour pressure deficit, precipitation etc) which can be more directly compared.*

However, as we are running under the ISIMIP 3b protocol, we cannot add an SSP2-4.5 run. ISIMIP uses 1-2.6, 3-7.0 and 5-8.5. As noted in the text, we chose not to use 5.8-5 due to well-trodden concerns around its plausibility.

2. The "direct human forcing" experiment design conflates multiple effects. The 2020 DHF runs hold cropland constant but allow natural vegetation to change transiently (lines 213–215). This is a reasonable pragmatic choice, but it means the DHF signal includes interactions between fixed human land use and dynamically changing natural vegetation. The paper does not adequately discuss how this confounding affects interpretation. A sensitivity test, or at least a thorough discussion of this issue, is needed.

Thank you for this. As we note in the methods, the models conceptualise land cover in ways that prevent direct comparison. Notably, WHAM-INFERNO has detailed representation of pastoral burning, whilst the Haas-model treats pasture as natural grassland. For the Haas model, where land covers outside of cropland are implicitly treated as natural, a simple sensitivity test will be conducted – ie. A run in which socio-economic and land cover forcing is all held constant.

However, we disagree that the interaction of vegetation composition and socio-economic change is necessarily confounding in the case of WHAM-INFERNO. The composition of natural vegetation <u>should</u> shape human decision-making around fire use and management. For example, increased wet tropical tree cover will (all else equal) make prescribed fire a less useful tool - as introducing disturbance reduces the natural fire resistance of the vegetation. This comes down to model conceptualisation – whether the socio-economic drivers of fire regimes are co-

responsive to ecological changes (WHAM-INFERNO) or treated as an exogenous forcing (Haas). In addition to the Haas sensitivity check, to clarify we propose to add the following text to the discussion at or around original line 512:

One way in which the models' contrasting structures manifest is in the treatment of land cover. At a simple level, the Haas model implicitly treats ecosystems outside of cropland as natural, whilst WHAM-INFERNO treats pastures and semi-natural grazed rangelands as distinct from unmanaged natural grasslands. However, whilst the Haas model conceptualises socio-economic factors as an exogenous forcing on burned area, WHAM-INFERNO assumes that human fire use and management is driven by the dynamic interplay of socio-economic and ecological factors – including vegetation composition. WHAM AFTs' fire use for fuel-load reduction dynamically adjusts to tree cover, for example. This reflects differing system conceptualisation – but also makes isolation and comparison of modelled socio-economic forcing necessarily imperfect.

3. No formal uncertainty quantification or statistical testing of inter-model differences. The paper's central claim — that inter-model disagreement on direct human forcing exceeds inter-scenario disagreement — is supported only by comparing Pearson correlation coefficients (e.g., r = 0.35 between models vs. r = 0.52 within WHAM-INFERNO across SSPs; lines 357–360). No confidence intervals, significance tests, or formal uncertainty decomposition are provided. Given that this is the paper's headline finding, more rigorous statistical support is essential.

Thank you. We accept that the study's findings would benefit from additional statistical support. We propose the following:

- *A pairwise factorial analysis of the effects of the model structure (Haas vs WHAM-INFERNO), scenario (SSP1.26 vs SSP3.70) and ESM driving data (GFDL, MRI, UKESM).*
- *This will be based on pairwise differences in burned area averaged across 2020-2025 and 2095-2100.*
- *We will use spearman's rho and mean absolute difference to quantify the impact of these factors on model sensitivity to climate and socio-economic forcing. On reflection, the zero-inflated data make rho more appropriate than Pearson's r even with a square-root transformation of data.*
- *We will use bootstrapping to calculate confidence intervals; this will be done by randomly sampling with replacement from the GFED regions.*

Minor Comments

4. Lines 341–348: The IAV comparison finds that WHAM-INFERNO underestimates and the Haas model overestimates IAV relative to GFED5. However, the Haas model's IAV is substantially higher in SSP1 (61.8 Mha) than in SSP3 (40.8 Mha), which is counterintuitive — shouldn't stronger climate forcing produce higher IAV? This deserves explanation.

Yes, this is because of the process of detrending used in the original manuscript. Ie the Haas SSP3 run had a strong increasing trend, whilst the SSP1 run had little trend and hence more random noise as a proportion of overall changes. We will use factorial runs to address the quantification of IAV more completely. IE, by conducting climate 2020 runs, we will be able to directly measure the IAV, rather than the detrending method used previously.

5. Line 171: Typo — "Mangeon et al., 20216" should be "Mangeon et al., 2016."

6. Lines 475–483: The discussion of anthropogenic fragmentation is framed almost entirely in terms of fire spread mechanics. However, the motivation for better representing fragmentation in fire models should also be grounded in its demonstrated ecological importance — for example, anthropogenic fragmentation significantly impacts biodiversity (https://doi.org/10.1111/brv.12519) and animal behavioral adaptation (https://doi.org/10.1002/ece3.71721).

Helpful comments, thank you.

7. Lines 232–238: The comparison of IAV between model outputs for 2020–2029 and GFED5 observations for 2011–2020 is acknowledged as indirect, but the decade shift also implies different climate baselines. This issue should be stated more explicitly.

Please see above, we think the difference between a climate 2020 run and runs with transient climate will provide a more robust measure of IAV. We will be more explicit about the limitations of this indirect comparison.

Reviewer 2

In this manuscript, the authors perform a sort of miniature model intercomparison project, forcing the WHAM-INFERNO and Haas fire models with a standard set of inputs to explore how they respond to changing climatic and socioeconomic drivers. These two models fall on opposite ends of the "representation of human effects on fire" axis, the goal was to explore how those differences in model structure (among others) manifest in terms of burned area through 2100 under two different future scenarios.

Overall, this is a neat little paper. There's not a lot here, exactly, but I think it does contribute something important to the literature: Comparing models developed recently with the aim of addressing separate known issues in global fire modeling.

The manuscript is well-organized and the writing quality is good. However, there are a number of points where the authors make interpretive leaps that seem unsupported by the presented analyses, as well as under-explained analyses. I would like to see these addressed and have given more detail below.

Substantive comments

• L136-140: If the coupling is "offline," does that mean the fuel burning from WHAM fires doesn't impact INFERNO flammability? Wouldn't that cause the total WHAM-INFERNO burned area to be artificially high?

We adjust for this in the offline model setup. We describe this in Perkins et al., (2025). Specifically, we added an equation to account for this that linearly reduces burned area after a threshold. We chose this function shape as it approximates the theoretical change in burned area in response to vegetation fragmentation (Archibald et al., 2012) – in our case driven by previously burned patches.

We have now implemented an online coupling of WHAM with JULES-INFERNO, including the dynamic response of vegetation to fire. The burned area results are quite similar to those presented here. Overall, this process is accounted for and we are confident it does not confound our results. We propose the following text at lines 146.:

The coupling presented here is offline and does not mechanistically represent vegetation responses to fire. Therefore, as in Perkins et al., (2025a), we apply an empirical adjustment to account for intra-annual fire-induced vegetation mortality. This linearly reduces burned area in a pixel after a threshold value is reached. The threshold and rate of reduction were determined empirically through calibration.

• L164-166, Table 2: Why bias-correct climate forcings from ISIMIP3b, which are already bias-corrected?

Adjustments to climate forcings were conducted for harmonisation rather than bias correction in a strict sense. We did this because:

a) *The Haas model was parameterised with ERA5, and as an empirical model we did not want to introduce possible model artefacts due to differences in forcing data.*
b) *For variables in WHAM-INFERNO for which we used JULES outputs rather than bias-corrected inputs, we use historical ISIMIP3a outputs driven by ERA5 data as the baseline.*

Together, these two factors led us to decide to rebase the ISIMIP3b climate forcing explicitly to an ERA5 baseline.

• L220: "with model projections of DHF"? Shouldn't it be "model outputs" or even better "inter-model differences in outputs"? What is a model projection of DHF if DHFs are inputs? … Based on text later in the paper, I think it would be clearer throughout to refer to this as "socioeconomic effects" or similar. "Direct human forcing" in most of the literature refers to things like land use etc. that are model inputs.

Thanks for this. We were adopting the terminology used in Burton, Lampe et al., 2019 – who use the terminology of the "direct human forcing of burned area" in the FIREMIP ensemble. However, we agree this is confusing.

We will change the terminology to: (model sensitivity to) socio-economic drivers of burned area.

• L232-233: Why wasn't a direct evaluation of IAV possible? The possibility of improved IAV performance is important enough that I feel like an extra model run covering the years in question would be justified.

This paper is not focused on model evaluation, but rather the model sensitivity under differing possible futures. Both models have had their historical outputs evaluated in separate papers. Hence, additional historical evaluation of the Haas model is out of scope. However, we will include additional experiments that allow the IAV to be more precisely extracted from the model outputs – i.e. by running models with climate held constant at 2020 levels.

In addition, in the introduction we will introduce what prior historical evaluations revealed about the models' capacity to reproduce temporal trends and IAV (as presented in Haas et al., 2026 & Perkins et al., 2025). Intuitively, this is that WHAM-INFERNO better captures the historical decline in burned area, whilst the Haas

model better captures IAV. In the manuscript draft, this discussion was restricted to the finding that both models achieve very similar metrics regarding the spatial distribution of burned area. We propose discussing model evaluation more thoroughly after the differences in model structure are discussed. So lines 73-76 would be cut and replaced with the following after line 95:

In spite of these structural differences, historical evaluation of the models has shown the two models are approximately equally able to reproduce the observed spatial distribution of burned area (Haas et al., 2022; Perkins et al., 2025a) – and that they are better able to reproduce the observed spatial distribution of burned area than models that participated in the latest published FireMIP assessments (WHAM: r = 0.83, Haas: r = 0.84, FIREMIP: 0.50 < r <= 0.81; Teckentrup et al., 2019). However, WHAM-INFERNO underestimates the historical interannual variability (IAV) of burned area (standard deviation of detrended outputs = 9.5 Mha; Perkins et al., 2025a), whilst the Haas model has IAV closer to observations (16.0 Mha; GFED5 = 23.7 Mha; Haas et al., 2026). By contrast, in line with observations, WHAM-INFERNO reproduces a declining burned area trend (-9.2 Mha yr^{-1}; GFED5 = -12.8 Mha yr^{-1}), whilst the Haas model suggests a modest increasing trend (+7.9 Mha yr^{-1}; Haas et al., 2026). These contrasting historical trends are consistent with WHAM-INFERNO giving greater emphasis to socio-economic factors (capturing the declining burned area trend) and climate the Haas model giving greater emphasis to ecological factors (capturing an appropriate degree of IAV).

• Fig. 2 (L270):

o Is it right that the middle of the Amazon is projected to be burned with "vegetation" managed fire about every 10 years under both SSPs? That feels high! Which exactly of the "vegetation" management types is the cause of that? How does it compare to the present day?

o A supplemental map with land cover fractions would help the reader interpret these results.

o The color bar would really benefit from a limited number of bins; it's basically impossible to distinguish any color from about 0.1 to 0.2.

This is very helpful, thank you. We will try to make the figure clearer through binned ranges. This apparent burned area in the Amazon is indeed due to the LUH2 land cover data, which drives WHAM! In 2100, even in SSP1.26, the data project a large decrease in forest area in the Amazon. The consequence of this is that much of the current Amazon effectively becomes a savanna (see Figure R1).

This land is then occupied by the hunter-gatherer agent type by WHAM!, because there is little agriculture there in the LUH2 data, and population remains very low. The fire in WHAM! projections is therefore indigenous fire, used for hunting, gathering and fuel-load reduction and fragmentation. This method, so-called "patch burning", is widely used in savannas in South America and sub-Saharan Africa. We think this is coherent with a world in which the Amazon basin does turn to savanna, rather than permanent agriculture, but also acknowledge this is partly a model artefact due to incoherence in the respective input data.

In the dynamic WHAM-INFERNO model runs, we update the forest cover with JULES PFTs. Therefore, this fire use is reduced (as JULES PFTs have a lower rate of amazon forest cover loss than LUH2). In the initial version of the manuscript, we plotted the data using LUH2 (as WHAM by default uses LUH2 inputs for land cover). In revision, we will show the outputs with this dynamic land cover adjustment, which is more intuitive to the reader.

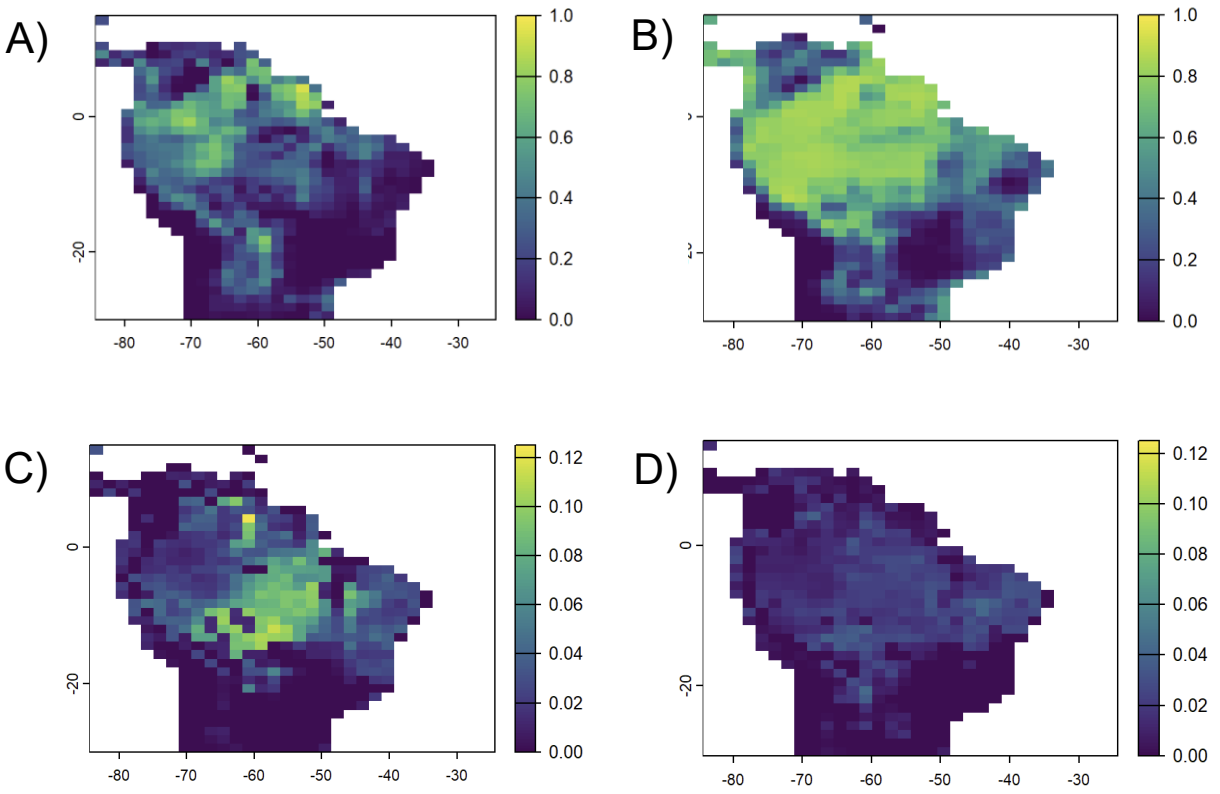

Figure R1: Comparison of forest cover (A, B) and indigenous fire use (C, D) in SSP1.26 in 2100. A) gives forest cover for LUH2, whilst B) gives JULES tree PFT coverage under UKESM forcing. C) gives indigenous burned area fraction given LUH2 tree cover, and D) gives the same adjusted for JULES PFT distribution. Colour schemes give pixel fractions of the respective commodity.

• L310-313: There doesn't appear to be evidence for this interpretation of climate vs. socioeconomic drivers in Fig. 6.

This section will be rewritten as we will now compare runs with climate held constant at 2020, as well as socio-economic factors held constant at 2020 (R1 major comment 1). This will allow us directly to diagnose the respective impact of different forcings.

• Fig. 7 (L325): This is another situation where a limited number of color bins would be helpful.

• L316, "where the climate signal is least disrupted by DHF in the present": Where is the evidence for this statement?

Thanks. We will change this text to:

"This includes regions with low population density and small agricultural footprints in the present-day, such as boreal North America and boreal Asia".

• L334-340: Figs. S1 and S2, the only ones referenced in this paragraph, don't seem to include these analyses—they're just maps. Scatter plots or something would be helpful.

We will change these maps into bar plots for easier inter-region comparison.

• L347-348: Again, I don't get the evidence here, although the statement is vague. Maybe just leave it out of the Results if it's actually explained in the Discussion.

This point will be made more explicitly in the new factorial experiments (i.e. by explicitly quantifying the role of biophysical vs socio-economic forcing). We will draw on these additional findings to make this point explicitly in the discussion section, as you suggest.

• L355-356: What is this analysis? Was it mentioned in the Methods? What exactly is being compared here, gridcells in 2100?

• L358-359: Fig. S3 does not appear to show anything about intra-model correlations across scenarios.

• L362-364: "there is greater agreement in determining how DHF will respond to socio-economic development (which dominates SSP1) than to a changing climate (which dominates SSP3)." This "dominates" language is strange. Socio-economic development and climate change are very different things, so it feels a bit comparing apples and oranges. Thinking strictly about influence on burned area and comparing Fig. 9 with Fig. 6 (a comparison this sentence should reference), it's not accurate to say that socio-economic development "dominates" SSP1 for Haas, although it does for WHAM.

• L 379-381: Where are these correlations coming from? Gridcell-by-gridcell where DHF has a positive effect?

Thank you for this set of comments. Overall, we think these issues, about a lack of clarity in the description of methods and a lack of precision in our interpretation of results echoes comments 2&3 of R1.

As noted above, we intend to address these comments by more clearly isolating the climate and socio-economic signals in the results and by revising the methods used to assess divergence (similarity) across model outputs. By stating this approach more clearly in the methods section, we can later avoid the issues you identify here.

Consequently, we expect the specific pieces of text you point to here most likely to be removed or at least very substantially revised in any revised versions of the paper.

Minor comments/corrections

• Throughout: Figures are too-resolution and some (e.g., Fig. 4) show JPEG artifacts. Per the GMD manuscript composition guidelines, JPEG should only be used for photographs. Maps etc. should be PNG format instead.

We will revise figures to be higher resolution and to avoid these issues.

• L20: "of" should be "on"

• L67 and throughout: "ensemble" doesn't seem like the right word for the coupled WHAM-INFERNO system. "Ensemble", at least in my experience, is about groups of model runs from either a single model (e.g., an Earth System Model might run 5 simulations with slightly varied initial conditions to sample over uncertainty) or different models (e.g., the FireMIP phase 1 "ensemble" of models).

Yes, model terminology can be very different across fields. In coupled human-natural systems, it would be normal to talk about a coupled ensemble as being a biophysical model and a socio-economic model coupling. To avoid this confusion, we will describe WHAM-INFERNO as "the combined WHAM-INFERNO model" or just "WHAM-INFERNO".

• L73-75: This seems like a result in the middle of the Introduction

Thanks, we will better explain that this refers to previous evaluation of the models, with references, rather than analysis conducted for this paper.

• L122: “indigeous" typo

• L209:

  o “minima” (plural) should be “minimum” (singular).

  o -425 does not look right for unmanaged fire

Yes, the -425 figure is the combined effect of unmanaged and managed fire, thank you for catching that. We will correct this.

• L421: Refer to Fig. 1.

• L436: Refer to Fig. 1.

• L553-555: Spaces missing between various pairs of words.

• Fig. S2: Mention that these are GFED regions.

We are happy to add these clarifications.

• Fig. S3: What is “anom.”? Is that supposed to be “change over 2020-2100”?

Thanks, this should be clarified, it is the change between 2020-2100.