Response to Reviewer #1

R1.0 This manuscript presents an interesting ensemble-based framework for characterizing extreme wildfire events using a process-based global fire model forced by large climate ensembles. The methodological setup is promising and timely. However, the current manuscript remains largely descriptive, and several methodological assumptions, statistical choices, and interpretations require stronger justification before the conclusions can meet the standard of a top-tier journal.

We thank the Reviewer for the overall assessment. Below we individually respond to the referee comments with our responses in blue. Specifically, we suggest revisions to the manuscript accordingly, strengthening the interpretation of our findings, providing stronger justification of methodological choices, and more clearly articulating the novelty and unique contributions of the study. The revised text is indicated in ***bold italic blue*** following the answer to each individual comment below.

R1.1 The central conclusion that a 1440-year pooled ensemble captures more extreme wildfire events than a 36-year single realization is largely expected from sampling theory. While important, this result alone offers limited novelty. The manuscript would be stronger if it provided deeper physical insight into what specifically drives the most extreme wildfire events.

Thank you for the comment. We believe that the central conclusions and novelty of the manuscript extend beyond the expected outcome from sampling theory, and would like to clarify the main conclusions of the manuscript and the physical insights each provides.

First, while it is expected from sampling theory that a larger sample such us SMILE-fire samples more extremes, we would like to emphasize the distinction between this expectation and the contribution of demonstrating it using a mechanistic process-based vegetation-fire model. To our knowledge, this is the first study to apply a bias-adjusted SMILE preserving modeled internal climate variability to a DGVM with vegetation-fire interactions (Lines 67-68). Quantifying how much the resulting SMILE-fire maxima exceeds single realizations maxima is itself an important contribution because it demonstrates that current state-of-the-art climate impact assessments with such DGVMs, which rely on single climate forcings, miss a substantial portion of plausible extreme wildfire events due to internal climate variability.

Secondly, this demonstration carries implications beyond the sampling effect itself, because we show that from the climate hazard to the impact level, the burned areas and fire carbon emissions are more underestimated by single realizations in contrast to the FDI large ensemble. In other words, propagating internal climate variability in LPJmL-SPITFIRE has a larger effect on burned area and emissions than on FDI, because of the fire-vegetation interactions. This is, to our knowledge, a finding that has not previously been demonstrated using a process-based large-ensemble framework before, since large ensemble fire studies focus largely on fire weather indices and do not represent this feedback mechanism.

Third, the manuscript provides additional physical insights that we should move beyond characterizing fire weather only, because the large ensemble results indicate that extreme FDI is a necessary but not sufficient condition for extreme BA and FireC. A conclusion that shorter simulations or observations would not be able to demonstrate so robustly and that can only be captured by a process-based fire model explicitly. We agree that disentangling the relative roles of climate drivers, ecosystem processes and non-climate drivers like human ignitions would provide additional physical insight, and we see this a valuable direction of future work and another manuscript in itself. In this respect, we note a recent study using LPJmL-SPITFIRE simulations for Europe using factorial experiments (Billing et al., 2026), which quantifies the effects of climate and non-climate drivers in future European burned area. The conclusions and added value of the present manuscript hold nonetheless. While Billing et al. (2026) isolates the effect of different drivers using single realizations, instead we use a SMILE and establish for the first time that more broadly representing internal climate variability in process-based modeling reveals a wider range of physically plausible extreme wildfires in terms of impacts such as burned areas and fire carbon emissions. We quantify this by how much across different fire regions, and provide evidence of the value of impact-centric approaches beyond fire weather.

We will revise the second paragraph of the discussion accordingly, strengthening the interpretation of the processes behind the larger differences between SMILE-fire and single-realization simulations in BA and FireC relative to FDI, and thus more clearly articulating the novelty of combining a climate large ensemble with a process-based vegetation-fire model to sample extreme wildfire impacts, rather than fire danger only:

'While it is expected from sampling theory that a larger sample such us SMILE-fire samples more extremes, the contribution of demonstrating this using a mechanistic process-based vegetation-fire model carries implications beyond the sampling effect itself. From the climate hazard to the impact level, the burned areas and fire carbon emissions are more underestimated by single realizations relative to the large ensemble than FDI is, *suggesting that the SMILE-fire simulations are particularly valuable for studying fire impact variables (burned area and fire carbon emissions) compared to the climate driver (fire danger). In other words, smaller changes in fire danger extremes due to internal variability have the potential to trigger stronger responses in burned area and emissions.* ***We interpret this as nonlinearities and vegetation-fire interactions that originate from the SMILE-LPJmL-SPITFIRE model chain. First, the input SMILE preserving internal climate variability determines the FDI together with dynamically simulated vegetation density. Then, the potential ignitions (natural and anthropogenic) are scaled by the FDI into successful ignitions, which spread using dynamically simulated fuel loads, and the resulting fuel consumption and vegetation mortality feed back into the vegetation component for subsequent fires (see Appendix B). Hence, while FDI mainly reflects initial condition differences among the climate ensemble members, burned area and emissions also depend on vegetation and fuel availability. Higher fire danger can increase the likelihood of successful ignitions and, once a fire is established, favour faster and more extensive spread under favourable fuel and weather conditions. Consequently, relatively small differences in fire danger can result in larger differences in the number and extent of fires, which then produce larger differences in fuel consumption and vegetation mortality.***

This same model chain effect also explains why the degree of undersampling differs among burned area and fire carbon emissions. *For most regions, the maxima of single realizations rank below those of SMILE-fire, with fire carbon emissions from single realizations showing the largest undersampling.* ***While burned areas mostly result from the Rothermel-based rate of spread, fire carbon emissions depend on the burned areas and additional modeling layers.*** *In our interpretation, fire carbon emissions are particularly more sensitive because they depend not only on burned area but also on fuel load and combustion completeness, and thus exhibit higher sensitivity to the role of internal climate variability on the process-based modeling chain. This interpretation is also consistent with the pairwise correlation structure among FDI, BA and FireC: the strongest relationship is between BA and FireC, with more modest relationships between FDI and BA or FireC, reflecting the additional modeling steps that accumulate along the model chain.'*

New reference:

Billing, M., Forrest, M., Bowring, S. P. K., Oberhagemann, L., Neidermeier, A. N., Hetzer, J., Bloh, W. Von, Müller, C., Rolinski, S., Hickler, T., & Thonicke, K. (2026). Future Projections of Burned Area in Europe Highlight the Importance of Human Action. https://doi.org/10.1111/gcb.71043

R1.2 The study treats 40 ensemble members as a single pooled 1440-year sample. This assumes the ensemble members are statistically exchangeable and sufficiently independent. The authors should better justify this assumption and discuss possible implications for extreme-value estimation.

As the reviewer notes, we indeed treat the 40 ensemble members as a single pooled 1440-year sample in Figures 3 and 4 to show that the tails of the distribution are better sampled by large ensembles. This is consistent with the design and standard usage of SMILEs, see e.g. Deser et al. (2020), Maher et al. (2021), Tebaldi et al. (2021), Bevacqua et al., (2023). By design, large ensemble simulations are generated by running the same climate model many times each with different initial conditions, representing independent simulations but under the same underlying climate (because it's the same model and scenario). As no member has special status over the others (the initial conditions are generated randomly and the members are all affected by the same forcing) the joint distribution of members does not change under permutations, i.e. the members are exchangeable. They are further conditionally independent, conditional on the shared forcing and climate model / DGVM, as they share no information beyond this. This allows pooling of the members to derive return periods, conditional on forcing and model.

In addition to pooling across ensemble members we also pool temporally, consistent with standard praxis in extreme value theory in which block maxima of sufficiently long blocks are treated as sufficiently independent. As transient simulations covering 1979-2014 with increasing anthropogenic and natural forcings, a trend will exist in each ensemble member. However, we detrend the FDI, BA and FireC to enable comparisons regardless, as said in Line 48. This justifies the pooled samples to be considered independent realizations of internal variability in the case of Figures 3 and 4.

In terms of implications for extreme value estimation, the size of the ensemble can be a factor, as discussed by Tebaldi et al. (2021). Accordingly, our ensemble size is within the required range. However, we note that this answer may differ for impact metrics, i.e., the DGVM output and not the SMILE input, since the extremes in the manuscript concern the output of a DGVM rather than the climate forcing itself.

We will clarify the question of exchangeability, independence, and their implications for extreme-value estimation in the manuscript as outlined above. In section 2.3.1. we will add a paragraph in the end: ***To calculate the percentiles, we treat the 40 members as statistically exchangeable and sufficiently independent, consistent with the standard use of SMILEs in literature (Deser et al., 2020; Maher et al., 2021; Tebaldi et al., 2021; Bevacqua et al., 2023). This approach treats each member exchangeably, differing only in their initial conditions of the forcing SMILE. Although they share the same external forcing, we assume each member can still be considered sufficiently independent because we focus on detrended LPJmL-SPITFIRE output to calculate the percentiles. In terms of ensemble size, the 40 members are within the range identified for robust estimation of extreme values for climate variables using SMILEs (Tebaldi et al., 2021).***

New references:

Tebaldi, C., Dorheim, K., Wehner, M., & Leung, R. (2021). Extreme metrics from large ensembles: Investigating the effects of ensemble size on their estimates. Earth System Dynamics, 12(4), 1427–1501. https://doi.org/10.5194/esd-12-1427-2021

Maher, N., Milinski, S., & Ludwig, R. (2021). Large ensemble climate model simulations: Introduction, overview, and future prospects for utilising multiple types of large ensemble. Earth System Dynamics, 12(2), 401–418. https://doi.org/10.5194/esd-12-401-2021

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., & Ting, M. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. Nature Climate Change, 10(4), 277–286. https://doi.org/10.1038/s41558-020-0731-2

R1.3 Return periods are inferred directly from empirical percentiles (e.g., p99 as a 1-in-100-year event). This approach is somewhat simplistic for rare-event analysis. A formal extreme-value framework such as GEV or peaks-over-threshold would provide more robust tail characterization.

We would like to emphasize that a central advantage of using SMILEs is precisely that they provide a sufficiently large empirical sample to characterize extreme values without relying on parametric or tail extrapolation assumptions (i.e. that we are in the domain of attraction of a GEV) as in extreme value theory. Further, having a SMILE estimate allows to analyse the dynamics of the events, as the dependence between FDI and BA and FireC. We agree that a GEV or peaks-over-threshold approach would be a valuable addition to extrapolate beyond the range supported by our pooled sample size, however, in this work we decided to keep within that range.

We will add a sentence in this regard in the end of the section 2.3.1: ***A central advantage of using large ensembles is that they provide a sufficiently large empirical sample to characterize extreme values without relying on parametric assumptions.***

R1.4 The manuscript compares model output with GFED mainly using mean variability and correlations, but the study focuses specifically on extremes. Model skill in reproducing uppertail burned area and extreme fire-emission events should be more explicitly evaluated.

Thank you for the comment, this question was also raised by Reviewer #2 and we agree that this is a relevant point. In addition to Figure 2, the Appendix Figures C6, C7 and C8 illustrate the annual maximum values of FDI, BA and FireC at the pixel level across the different simulations. In particular, Appendix Figures C7 and C8, compare the burned areas and fire emissions simulated by LPJmL-SPITFIRE against GFEDv5.1. Both the maps and the distributions when pooling all pixel values of the temporal maxima (C7f and C8f) show (1) a good spatial agreement in the upper tail values and (2) that, as expected, SMILE-fire maxima values are consistently larger than those in single simulations.

We agree that this information is currently overlooked and propose to better highlight the ability of LPJmL-SPITFIRE to reproduce the upper tail of the GFEDv5.1 BA and FireC distributions.

After line 219, section 3.1, we suggest adding a new last paragraph: ***The ability of LPJmL-SPITFIRE to reproduce the spatial variability of the GFEDv5.1 BA and FireC extremes is shown in Appendix Figures C7 and C8. In general, a good spatial agreement is found at the pixel level. Although BA and FireC show highly skewed distributions (Figures C7e and c8e), reanalysis-fire and GFEDv5.1 agree closely in the log-transformations when pooling all the maxima values of all pixels (Figures C7f and c8f), and SMILE-fire shows a heavier upper tail consistent with its broader sampling of internal climate variability.***

In addition, we will revise the discussion paragraph that addresses the SPITFIRE challenges in modelling extremes: ***'While LPJmL-SPITFIRE produces reasonable estimates of mean and extreme burned area and fire carbon emissions, several aspects of extreme fire behavior remain less explored.*** *In particular, SPITFIRE represents surface fire spread but does not explicitly simulate crown fire or canopy fire spread, which are characteristic of the most extreme fire events in boreal and other fire-prone ecosystems (Ward et al., 2018). Additionally,* ***as the live fuel moisture levels are calculated based on a growing season index (Oberhagemann et al., 2025)****, including the influence of wind and precipitation on live fuel moisture computation would further contribute to improve extreme fire danger conditions and to separate the influence of wind speed on fire spread from the influence of atmospheric drying of the fuel. (...)'*

R1.5 Using a VPD-based fire danger index as the primary climate driver may be overly simplistic. Wildfire extremes are often controlled by compound hot, dry, and windy conditions, along with antecedent fuel moisture. The manuscript should better justify this choice and discuss its limitations.

Thank you for the comment, a similar question was also raised by Reviewer #2 and we agree that the manuscript can benefit from more clarity about the implications of the VPD-based FDI. We would like to clarify that wind and fuel-related processes, which the reviewer notes are not part of the FDI itself, are represented explicitly elsewhere in the LPJmL-SPITFIRE model chain, after the calculation of the FDI.

The VPD-based FDI was introduced into LPJmL-SPITFIRE by Drüke et al. (2019), replacing the previously used Nesterov index. The performance of the model significantly improved in terms of burned area in South America by integrating the air humidity, precipitation and vegetation via the VPD-based FDI. Additionally, the FDI by LPJmL-SPITFIRE is not the sole determinant of fire behavior. The spread of the fires started by ignitions scaled by FDI is subsequently determined by the wind-driven

Rothermel model, using fuel loads provided dynamically by the LPJmL vegetation component. In this sense, wind and fuel load are represented mechanistically later in the model chain.

We will improve the description of the different modelling steps of LPJmL-SPITFIRE in Appendix B and also in the discussion, where we will comment on the limitations associated with VPD-based FDI as the single climate driver in our analysis regarding driver-impact relationships.

Appendix B: *'SPITFIRE considers two types of ignition sources: lightning flashes and human ignitions determined based on population density (see Eqs. 3 and 4 in Thonicke et al. 2001).* ***Following a potential ignition event, the number of ignitions is scaled using a fire danger index (FDI), representing the proportion of ignitions that successfully become spreading fires.*** *While previous versions of SPITFIRE used the Nesterov Index (based on maximum temperature, dew point temperature and the different plant functional types) for calculating fire danger, recently Druke et al. (2019) implemented a VPD-based FDI.* ***The formulation of the VPD-based FDI by Druke et al. (2019) combines air temperature, relative humidity, precipitation, and vegetation density dynamically simulated by LPJmL.*** *The duration of the fire is determined by the FDI and assumes a longer burn time under high fire danger. The Rothermel equations are employed to determine wind-driven fire spread, which also varies according to plant functional types. Finally, the model uses the simulated fire danger, ignitions, and spread to estimate burned area and fire carbon emissions. The model restricts the simulation of wildfires occurring to natural vegetation, excluding managed land, such as agriculture, due to fundamentally different fire dynamics in this type of land (Ribeiro et al. 2024).* ***As shown in Fig. B1, the main components of SPITFIRE consist of (1) ignition module, (2) fire danger module (3) fire spread module, and finally (4) fire impacts (burned area, fire carbon emissions and vegetation mortality). Together, these 4 components determine the successful ignitions (fire occurrence), the fire danger (intensity and duration), fire spread (wind-driven Rothermel model) and the resulting impacts (burned area, emissions) which later feedback into the vegetation component of LPJmL for subsequent time steps.'***

Discussion: *'There are some limitations of the present study to note.* ***In contrast to older versions of LPJmL-SPITFIRE which used the Nesterov Index to represent fire danger, the current version of FDI is based on VPD, precipitation and plant functional types (PFT)-specific tuning parameters. Despite being one of the primary controls of fire, we acknowledge that the VPD-based FDI does not explicitly account for wind, because in LPJmL-SPITFIRE the fire spread is calculated after the FDI via the Rothermel model.*** *Thus, while LPJmL-SPITFIRE framework does account for meteorological variables beyond VPD, their contributions are not explicit in what we refer here is as 'climate driver', as we define it through FDI (VPD-based). Disentangling the relative roles of other climate drivers beyond FDI constitutes an important avenue for future work. Particularly, assessing the ensemble-based relationships between climate drivers and extreme impacts (burned area and carbon emissions) could be generalized to a broader set of climate variables beyond FDI, including wind speed,* ***soil moisture****, and large-scale atmospheric circulation patterns. Such an approach would allow exploring physically self-consistent combinations of* ***compounding*** *drivers governing extreme fires, in line with physical climate storyline methods.'*

R1.6 Lightning and human ignition assumptions are largely fixed across ensemble members. This may suppress an important source of climate-driven wildfire variability, since ignition probability itself can vary substantially with atmospheric conditions.

Thank you for the comment, a similar question was also raised by Reviewer #2 concerning the lightning ignitions (please see full response to R2.3). While the raw lightning and anthropogenic ignition potential is indeed fixed across members, both natural and anthropogenic ignitions are scaled by FDI before becoming successful ignitions, so some climate variability is indirectly represented in the ensemble members. As we note in R2.3 response, we acknowledge that this is a simplification in the methodological set-up: ideally a fully dynamically consistent design would vary lightning occurrence with each ensemble member meteorology. However, lightning data are not readily available as standard climate model output, let along for the large ensemble, and this setting is already novel in contrast to standard process-based modeling.

We will revise this point in the discussion accordingly: *'Beyond hydrometeorological drivers, topography and fire ignition assumptions are aspects that future work could further address. Satellite-derived lightning products have inherently limitations (Christian et al., 2003) and human-caused ignition based on population density (Thonicke et al., 2010) does not fully capture human ignition behavior driven by, for example, land management practices or socioeconomic factors (Ribeiro et al., 2024; Forrest et al., 2024). Furthermore, lightning ignition probability is held constant across input ensemble members, and thus does not respond to changes in convective activity associated with different climate states.* ***Nevertheless, because these raw ignition potentials are subsequently scaled by the FDI before becoming successful ignitions (see Appendix B1), successful ignitions do inherit some variability from each ensemble member through this scaling step, even though the underlying ignition potential does not. As a result, the simulated extremes in this study are driven primarily by the variability in FDI and the vegetation-fire feedback loops, rather than by variability in ignition occurrence itself. Therefore,*** *while fire weather favorable conditions across the ensemble members amplify the fire outcomes of the same prescribed ignitions, lightning cannot itself act as a source of climate-driven variability in the simulations.* ***Hence, in ignition-limited regions, our experimental design may therefore underrepresent the full extreme range that could arise if lightning variability were also allowed to respond dynamically to the simulated climate of each ensemble member.*** *Future work could include climate model output of lightning-related variables to allow ignition probability to respond dynamically to the climate state, enabling a more complete representation of natural climate variability in fire activity (Janssen et al., 2023).* ***However, this is usually not standard output in SMILEs.'***

R1.7 The conditional metric (Δ cond) based on minima of conditional and unconditional distributions is not intuitive and lacks standard statistical interpretation. More conventional dependence metrics could improve clarity and robustness.

Thank you for this comment. We would like to emphasize that we also report a more standard dependence metric in Figure 2d-f, where we show pairwise linear correlations between FDI, BA and FireC, and in which correlations between BA and FireC show the strongest correlation, followed by FDI vs FireC and FDI vs BA. We note in Lines 201-210 that this hierarchy of correlations aligns with the LPJmL-SPITFIRE structure and that fire danger is not the solely determinant of burned areas and emissions.

The metric Δ cond is intended to complement the correlation analysis because it specifically captures the co-occurrence of extremes, and it quantifies how much lower FDI is than the maximum of FDI when BA and FireC are maximum. As stated in the manuscript in Lines 189-291, *While values near zero indicate that extremes in both variables tend to co-occur, negative values indicate that FDImax does not necessarily coincide with BAmax and FireCmax.* We propose to keep both the correlation analysis and the conditional metric Δ cond.

R1.8 The authors attribute fire danger–impact decoupling to fuels, ignitions, and vegetation feedbacks, but these mechanisms are not quantitatively decomposed. As a result, the mechanistic interpretation remains largely speculative.

Response: Thank you for this comment. In the manuscript this interpretation is supported by (a) the structure of LPJmL-SPITFIRE, (b) the correlation patterns (Figure 2d–f), (c) larger differences between SMILE-fire and single-realization simulations in BA and FireC compared to FDI (Figures 3 and 4) and (d) the Δ cond analysis (Figures 5 and 6). We agree the manuscript would benefit from improving this interpretation, which we aim to improve in alignment with the response to R1.1, R1.5 and R2.1.

In R1.1 we note that disentangling the relative roles of climate drivers, ecosystem processes and non-climate drivers like human ignitions would provide additional physical insight, and we note that this was part of a recent study by Billing et al. (2026) for Europe. In that paper, the authors identify fire weather and socio-economic development as the two dominant drivers of future European fire activity, with vegetation and land-use changes playing a secondary but regionally important role, especially when fire weather conditions are relatively mild. This also agrees with our interpretation, and we will add this reference to the introduction and discussion to support our conclusion of fire weather as a dominant but not exclusive driver of fire impacts.

R1.9 The analysis uses only one climate model and one fire model. Therefore, the study assesses internal variability but not structural uncertainty, which may be equally or more important for wildfire projections.

Thank you for this comment. We agree that we sample internal climate variability, not structural model uncertainty, and agree that it is important to be clearer on this distinction.

We will adjust the discussion in the lines 361-373 accordingly: *'While climate models and vegetation-fire process-based models are powerful tools to understand climate-vegetation-fire relationships, they are computationally expensive at the global scale with large ensembles.* ***In addition to uncertainties from internal variability, other sources of uncertainty arise from structure and parameters of the climate and vegetation models themselves, which can be approximated with multi-model ensembles.*** *Frameworks such as ISIMIP use data from multiple climate models and multiple impact models,* ***so they characterize structural (model) uncertainty, albeit they do this for a single climate realization per climate model. In that way, they typically underestimate uncertainties related to internal climate variability, and the structural uncertainty they do characterize can also be underestimated in key impact sectors, as only a modest selection of models are typically used (Bevacqua et al., 2026).*** *While our analysis consists of one SMILE paired with one impact model, a single SMILE provides multiple realizations of internal variability as represented by a single climate model. Combining multiple SMILEs with multiple process-based vegetation-fire models would be the combination of both approaches, further enabling explicit separation of the contribution of internal variability from structural model uncertainty (Lehner, 2024). Extended to future fire projections, a key area of study for the fire modeling community with high societal relevance (Keeping et al., 2025; Quilcaille et al., 2023, Billing et al. 2026), would additionally need to account for varying socio-economic scenarios (Lehner, 2024), which can be highly relevant for fire ignition and suppression. This remains computationally and methodologically challenging. These barriers are important to overcome for concrete wildfire risk and impact quantification. We here illustrate the added value of using large ensemble climate data, which might be the first step towards such multi-SMILE multi-impact model multi-scenario studies in future.'*

R1.10 Regions such as Europe and Boreal Asia show strong under sampling of extremes, but the manuscript provides limited physical explanation for why these regions behave differently. Stronger regional interpretation would improve scientific value.

In Lines 296-298 we note that EURO and BOAS are also among the regions with the greatest discordance between its maximum fire danger and its fire danger values during maximum impacts, and between its maximum burned area and its values of burned area when fire danger is maxima. In our interpretation, this helps explain why these regions show the strongest undersampling of extremes by single realizations relative to SMILE-fire. In other words, in regions where the more extreme fire outcomes depend more on factors beyond fire danger, the SMILE-fire is able to sample a much wider range of the combined climate and vegetation trajectories necessary to give rise to such extremes, something a single realization is less likely to sample comprehensively. We will revise the last paragraph of section 3.3 to make this connection more explicit:

'EURO and BOAS were among the regions where single realizations undersample extremes the most and they show more discordance between conditional and unconditional maxima, highlighting the critical role of other factors rather than climate conditions conducive of fire extremes. *This is consistent with the ability of SMILE-fire to sample a much wider range of the combined climate and vegetation trajectories that can give rise to such extremes, whereas a single realization is less likely to capture these same combinations comprehensively. These findings* ***also*** *underscore the critical role of ecosystem processes in modulating fire outcomes, which are explicitly captured in LPJmL-SPITFIRE. On the one hand, studies based on fire weather only lack the mechanistic representation of vegetation–fire interactions needed to explain why extreme fire danger does not always lead to extreme impacts. On the other hand, relying on a single realization, even with a process-based model, underestimates extremes due to insufficient sampling of internal climate variability. By* *combining a fire-enabled DGVM with a large ensemble, both the broad range of climate variability and the underlying ecosystem controls on fire can be robustly represented.'*

R1.11 All variables are linearly detrended prior to analysis. This may remove physically meaningful climate signals relevant to wildfire extremes. The rationale for detrending should be better justified.

We would like to clarify that detrending is applied to the output of LPJmL-SPITFIRE (FDI, BA, and FireC) and not to the climate forcing itself. The climate input to the LPJmL-SPITFIRE is not detrended.

We detrend the LPJmL-SPITFIRE output to enable comparisons between different data independent of trends when comparing LPJmL-SPITFIRE burned area and emissions with GFEDv5.1 and SMILE-fire and renalysis-fire. When comparing with GFEDv5.1, detrending is required because LPJmL-SPITFIRE does not capture the declining trend in burned area as illustrated in Appendix Figure C2. For consistency, comparing the reanalysis-fire, single-fire, and SMILE-fire simulations against each other detrending isolates internal climate variability (see reply R1.2).

Additionally, we also would like to clarify that the ensemble-based extreme driver-impact relationships analysis (Figures 5 and 6) is not detrended. Figures 5 and 6 caption do not mention detrending because in this analysis no comparsion between GFEDv5.1 or single-realizations is being made, only SMILE-fire is used.

We agree that this might not be clear in the manuscript and will revise Line 147 accordingly: *'Given that the LPJmL-SPITFIRE does not reproduce the observed human-induced declining trend in burned area (Andela et al., 2017; Oberhagemann et al., 2025) all* ***output data (not the forcing input)*** *were linearly detrended to enable comparisons independent of trends (Figure C2).*

R1.12 The manuscript frequently refers to extreme wildfire events, but a precise operational definition is lacking. It should be clearly stated whether extremes are defined by annual maximum burned area, emissions, percentile thresholds, or another metric.

Thank you for this comment. Throughout the manuscript, "extreme" refers to the highest value within a given sample. What differs between a given sample and analysis is the research question being addressed (Lines 72-76). As the reviewer notes, in the manuscript we focus on annual values, where burned area and fire carbon emissions were summed over each calendar year and GFED region, and the annual maximum of fire danger was obtained for each GFED region (Figure 2, Lines 142-145). In terms of research questions, first we address to what extent do SMILE-fire capture more extremes than single realizations (single-fire and reanalysis-fire), and secondly how the most extreme SMILE-fire fire danger relate to most extreme impacts. When comparing large ensembles with single realizations (Figures 3 and 4), we consider 'extreme' as the maximum value each different simulation (pooled SMILE-fire, single-fire and reanalysis-fire) is able to sample, and as a sensitivity analysis we compare with the p95, p97.5, and p99 percentile thresholds of the SMILE-fire (also extremes). When using the SMILE-fire to assess the relationship between extreme FDI, BA and FireC (Figures 5 and 6), we consider 'extreme' as the maximum of each individual ensemble member to assess whether extremes in different variables tend to co-occur in the SMILE-fire. In both research questions/analyses "extreme" refers to the highest value within each sample but with different designs tailored to the research question.

We suggest adding a clarification sentence in section 2.3 before 2.3.1 and 2.3.2: ***To address to what extent SMILE-fire captures more extremes than single realizations (section 2.3.1) and how the extreme SMILE-fire danger relate to extreme impacts (section 2.3.2), hereafter we refer to 'extreme' as the highest value within a given sample tailored to each research question. Below, section 2.3.1 and 2.3.2 detail the different analyses.***

R1.13 Figure 4 caption appears incorrect (Same as Figure 4) and should be corrected.

Thank you, we will update the caption accordingly.

R1.14 Figures 3–4 are visually dense and difficult to interpret at publication scale.

We propose making changes to improve the readability of Figures 3 and 4, including simplifying the color scheme and adjusting the y-axis of the boxplots (panels d-f), switching the map projection to Robinson (following a related suggestion by Reviewer #2), removing the meridian and parallel grid lines and repositioning the regional labels (panels a-c).

R1.15 Several equations and symbols are inconsistently formatted.

We will review the mathematical notations of the manuscript carefully.

R1.16 Language editing would improve clarity and readability.

We will review the text in order to improve readability.