

Reviewer 1

We thank the reviewer for their time and thorough evaluation of our manuscript. We believe that we have substantially improved the manuscript in response to the many suggestions provided. Given the extent of the revisions, we have not reproduced every change here; please refer to the tracked-changes version of the manuscript for full details.

General comments

This is another global and regional assessment study of changes in fire weather (FW) carried out based on: (i) a meteorological fire danger index calculated with values of meteorological variables/climate elements, such as precipitation, temperature and air humidity; and (ii) climate model simulations for different GWLs and scenarios (in this case, the surpassed RCP). Unsurprisingly, the study concludes that the values of the chosen index and metrics will increase and that the increases are proportional to the increasing severity of the GWLs and scenarios. This is one of the main problems with the manuscript: what's new?

The novelty of our study lies in inverting the usual framing: rather than focusing solely on how much fire weather will increase, we ask what this means for when and how fire management can be carried out. Specifically, we highlight shifts in seasonal timing, such as earlier transitions from High to Very High fire danger, and the persistence of low-fire windows that remain crucial for controlled burns and other preparations. This management-focused lens, centred on the timing of opportunities and risks, has not been systematically explored in global or regional fire-weather studies. In addition, our work (i) uses a perturbed physics ensemble to quantify uncertainty more comprehensively than typical CMIP-based approaches, and examine its implications, (ii) frames results in terms of Global Warming Levels (GWLs) including 1.5C and 2C as used in the Paris Agreement to enhance policy relevance, (iii) uses more than one emissions scenario for each WL to further extend the quantification of uncertainty, and (iv) integrates global patterns with regionally tailored insights for Australia, Brazil, and the USA. Together, these elements move the paper beyond documenting change, toward informing both mitigation benefits and adaptation strategies.

Having said that, we would like to note that we respectfully disagree with the premise that our study would not be of value if the findings merely agreed with those of previous work. Replication is an exceptionally important aspect of science, and for future climate change studies in particular, key questions absolutely must be addressed independently by different research teams using a variety of methods and models to build confidence. This is crucial when science is being used to inform decision-making or policy. For example, the assessment reports of the Intergovernmental Panel on Climate Change (IPCC) rely heavily on multiple lines of evidence in order to establish confidence, and this is particularly important for future projections. Therefore, even if the results of our study were not "surprising", it would still be important to publish them.

Additionally, this study suffers from other serious problems, which I will detail in the more specific comments. Together, these issues prevent you from recommending this manuscript for publication.

Specific comments

- As mentioned in the general comments, the first question that authors have to clarify and answer is what is new about their results/conclusions.

We have revised the abstract and discussion to more clearly articulate the novel contributions of this study. In particular, we now emphasise (i) the use of uncertainty estimates derived from a large perturbed physics ensemble to assess confidence in future fire weather projections across mitigation pathways, and (ii) the extension of fire weather analysis beyond peak season to the full annual cycle, explicitly linking projected changes to the timing and feasibility of Integrated Fire Management activities.

New abstract:

“Understanding future shifts in fire weather risk across peak-season, transitional, and off-season periods is crucial for adapting fire preparation and management to climate change. Fire management planning depends not only on how much fire risk is reduced through climate change mitigation, but also on how residual risk evolves under different warming pathways, including at low levels of global warming. Additionally, while most fire-weather projections focus on peak-season severity and length, fire management decisions - including prevention, preparedness, and controlled burning- are made throughout the annual cycle. This creates a growing information gap between climate-driven shifts in fire risk under different mitigation scenarios and year-round fire management frameworks. To address this, we explore future climate-driven fire-weather projections using the McArthur Forest Fire Danger Index (FFDI) and a large perturbed-physics ensemble, enabling a systematic assessment of uncertainty and confidence in projected changes globally and across three focus regions: Australia, Brazil, and the United States of America. We evaluate future fire weather across all phases of the annual cycle relative to both a historical baseline (1986–2005) and a recent reference period (2004–2023), under three Global Warming Levels (1.5 °C, 2.0 °C, and 4.0 °C) and two emissions scenarios (RCP2.6 and RCP8.5). In addition to changes in season length and peak FFDI, we quantify transitions between meteorological fire danger periods and shifts in low fire weather windows, linking projected climate change impacts directly to the timing and feasibility of Integrated Fire Management (IFM) activities. We project a global rise in FW at all GWLs, with the largest increases in Australia, then Brazil and the USA. At 1.5°C, the area exposed to Very High FW (FFDI ≥ 24) globally expands by 31% (25%–36%) relative to 1986-2005. Higher GWLs drive further increases, with more than a threefold rise in days with Very High FW from 2.0°C to 4.0°C. The transition from preparation periods to fire season advances by 7-36 days (Australia), 12–32 days (Brazil), and 5–36 days (the United States of America) at 2.0°C. Low FW windows persist, providing crucial opportunities for out-of-season preparation, though they become narrower with warming. Our findings highlight the dual need for mitigation and adaptive strategies, including accounting for changes in out-of-season fire risks. We offer an initial step toward a more dynamic form of IFM by illustrating how climate projections, impact metrics and seasonal diagnostics can be combined to inform preparedness, flexible planning, and providing a foundation for operational dynamic IFM.”

We have added a paragraph to the start of the discussion to make the novelty of this study more clear:

“Our model projections show that the more global warming we experience in the future, the higher the increase in peak fire danger, and the longer Very High fire weather seasons will become, extending over much more land area, especially at 4°C of warming. This supports evidence from previous studies (e.g. Jones et al (2022), Peng et al (2023)). Though our analysis providing a complementary line of evidence through a structured assessment of uncertainty across warming levels, emissions scenarios, and model parameterisations,

thereby strengthening confidence in the robustness of conclusions from these previous studies. Additionally, unlike previous studies, which mostly quantify the magnitude of these increases, our analysis focuses on what these changes mean for fire management, particularly the seasonal timing of preparation windows, controlled burns, and low-fire periods that remain critical for adaptation. This management-oriented framing inverts the usual problem: instead of only asking “how much worse will fire weather get” (Section 4.1), we also ask, “how and when can societies still act within an increasingly altered seasonal cycle?” (Section 4.2).”

- The second question is how much confidence the reader can have in the results and conclusions of the study. This question has to do with the methodological approach and methodology adopted. Studies on the impacts of climate change caused by global warming necessarily include two parts: (i) assessment of changes (e.g., through anomalies or ratios); and (ii) assessment of the statistical significance of these changes. This study only performs part of the assessment of changes, namely in terms of anomalies. Thus, the assessment of statistical significance, which is as fundamental as the first, is missing.

We agree that confidence in projected changes is central to the interpretation of climate change impacts. We note, however, that the original manuscript was not devoid of uncertainty or robustness assessment as the reviewer suggests. Throughout the initial submission, confidence was quantified using ensemble-based approaches that are standard and appropriate for perturbed physics ensembles, including ensemble agreement metrics, percentile ranges (e.g. 10th–90th and interquartile ranges), and the position of individual ensemble members within the full ensemble distribution. These approaches explicitly characterise uncertainty without relying on assumptions of normality and are widely used to assess robustness in ensemble climate projections.

That said, we recognise that readers may also expect explicit hypothesis testing when comparing distributions across baseline and future warming levels. In response to this comment, we have substantially expanded the use of formal statistical significance testing throughout the manuscript, complementing (rather than replacing) the ensemble-based confidence measures already presented.

First, for changes in peak fire weather at each Global Warming Level (GWL), we now test whether the distribution of FFDI values differs from the baseline period using a paired, one-sided Wilcoxon signed-rank test, with pairing by ensemble member. This directly assesses whether fire weather intensifies relative to the baseline for each emissions scenario and warming level. For all regions and GWLs examined, these changes are statistically significant ($p < 0.001$). While parametric t-tests yield comparable conclusions, we present Wilcoxon results due to the non-normal nature of the FFDI distributions.

Second, for the assessment of mitigation potential, we apply the same paired Wilcoxon framework to compare projected fire weather under mitigation scenarios against the high-end RCP8.5, 4.0 °C warming distribution. This allows us to formally test whether mitigation produces statistically distinguishable reductions in fire weather severity, beyond differences in anomaly magnitude alone.

Third, for the seasonal evolution of fire weather, we now include statistical testing across the full annual cycle. Changes relative to the baseline are assessed using a one-week running mean of FFDI, with significance evaluated using a paired, one-sided Wilcoxon signed-rank

test (paired by ensemble member). Significance levels are explicitly indicated in the figures, allowing robust seasonal shifts to be distinguished from changes within ensemble variability.

Fourth, for metrics related to fire management timing and prescribed burning windows, we complement formal testing with a likelihood-based ensemble agreement measure, reporting the percentage of ensemble members that agree on the direction of change. This approach is particularly appropriate for ensemble projections of threshold-based timing metrics, where directional robustness is often more informative than a single test statistic.

Finally, for driver attribution, we now apply a bootstrapping approach to assess the robustness of variable importance estimates, enabling inference based on distributional overlap rather than point estimates alone.

Taken together, the revised manuscript now combines ensemble-based confidence measures (percentile ranges and agreement metrics) with formal, non-parametric statistical tests wherever distributions are compared. This ensures that all major conclusions are supported by both physically interpretable ensemble robustness and explicit statistical significance testing, substantially strengthening confidence in the results and conclusions.

- The confidence in the findings is also related to the methodology adopted. In this sense, the authors need to answer the following questions:
 1. Why do you think an index developed for Australia, where wildfires are usually named bushfires, because of the main type of vegetation affected by the wildfires, is suitable for assessing fire danger anywhere in the world?

We acknowledge the reviewer's concern that the McArthur Forest Fire Danger Index (FFDI) was originally developed in Australia, where wildfires are often referred to as bushfires due to the colloquial term of dominant vegetation types rather than the type of fire. However, the FFDI itself is not vegetation-specific in its formulation; it is based on meteorological inputs—temperature, humidity, wind speed, and drought factor—that are fundamental drivers of fire danger across fire-prone ecosystems in many regions, not only in Australia. Previous studies have shown that the FFDI performs comparably to other indices, such as the Canadian Fire Weather Index, when applied outside its region of origin (e.g. Hoffmann et al., 2003; Dowdy et al., 2009). Also other studies have used it to assess global changes in meteorological fire danger under climate change, e.g. Peng et al (2023) . Most of these were discussed in the original m/s, and we have added remaining to the discussion (see reviewer 1s comment lower down on the discussion).

Our study design also reflects this consideration. While we present a global overview, our regional analyses focus on Australia, Brazil, and the western USA – ecosystems where the FFDI is particularly applicable due to biome similarities and fuel-limited fire regimes. We do not focus the conclusions on changes in the index to high-latitude boreal forests, where alternative indices may indeed be more appropriate.

In addition, the FFDI provides well-defined thresholds (e.g. “very high”) that are useful for exploring seasonal timing, preparedness, and prescribed burning opportunities – the central focus of our study. Applying a single index consistently across regions also allows for comparability in a global analysis.

We agree that a multi-index approach, or new developments in fire danger assessment (e.g. McNorton et al., 2025), would be highly valuable in future work, particularly in the context of multi-model ensemble studies. We see our analysis as a step toward that broader effort. See this adjusted part of the discussion:

“While the FFDI provides strong sensitivity to temperature and humidity in fuel-abundant ecosystems and has proven very useful for our temperate and tropical focus regions, using multiple fire danger indices in parallel could improve understanding and applicability across biomes. For example, boreal forests may benefit from indices better suited to high-latitude ecosystems, such as the Canadian Fire Weather Index (FWI), while other metrics like the Keetch–Byram Drought Index (KBDI) or the US Energy Release Component (ERC) capture additional aspects of fire danger relevant in specific regions. Applying a suite of indices would help build a more complete picture of future fire risk and reduce reliance on any single metric, representing an important avenue for future research, particularly for studies seeking global coverage. Beyond these established measures, there is also growing interest in machine learning and AI-based indices (e.g. McNorton et al., 2025), which can integrate diverse meteorological and ecological variables and may improve predictive skill in data-rich regions. Future assessments that benchmark traditional and ML-derived indices across regions could offer valuable insights into the strengths and limitations of each approach, and guide the development of hybrid frameworks that combine physical process understanding with data-driven methods.”

2. The suitability of FFDI for studying the fire weather globally can be measured based on the number of countries adopting FFDI as a fire danger indicator. So, it is important to know, how many countries use FFDI?

We appreciate the reviewer’s suggestion to consider the operational adoption of FFDI as a measure of its suitability. However, we would like to clarify that the number of countries using a given index in their operational fire danger rating systems is not necessarily a reflection of its scientific appropriateness for research applications. Fire agencies often adopt indices for reasons of institutional history, operational simplicity, or policy alignment, rather than because the index is uniquely well-suited in all ecosystems. For instance, the Canadian Fire Weather Index (FWI) is widely used operationally in Europe and North America, while the FFDI has been the standard in Australia, and has been adopted or adapted in South Africa and Spain.

In the scientific literature, however, the FFDI has been shown to perform well outside Australia, including in tropical ecosystems such as the Amazon (Hoffmann et al., 2003), and has been applied successfully in global fire weather studies (e.g., Golding and Betts, 2008; Jolly et al., 2015; Bett et al., 2020). Our study builds on this body of work and focuses specifically on ecosystems where the FFDI is particularly appropriate — temperate and tropical regions where fuel is abundant and meteorological conditions are the dominant drivers of fire occurrence and severity.

We agree with the reviewer that operational diversity is important and note in our future work section that multiple-index approaches, including the Canadian FWI and

newer machine-learning–based indices, would be valuable for expanding this type of analysis, particularly in ecosystems less well captured by FFDI, such as boreal forests and made in the discussion (see comment above)

3. Why do you think an index developed for grasslands is good for assessing meteorological fire danger in other types of vegetation?

We thank the reviewer for this comment, but would like to clarify a possible misunderstanding. The McArthur Forest Fire Danger Index (FFDI) was not developed for grasslands. It was originally designed in the 1960s by A. G. McArthur for use in eucalypt-dominated dry sclerophyll forests and woodlands in southeastern Australia (Noble et al., 1980). The index was empirically calibrated using observations of fire behaviour in these forest ecosystems, capturing the combined influence of key meteorological variables (temperature, relative humidity, wind speed) and antecedent dryness through the drought factor.

Australia also has a distinct McArthur Grassland Fire Danger Index (GFDI), which may be the index the reviewer is thinking of. GFDI was developed specifically for grassland fuels, highlighting how FFDI was intended for forested landscapes. In this study, FFDI is used as a meteorological fire weather metric to characterise atmospheric conditions conducive to fire, rather than to directly represent fire behaviour or impacts. Its strong sensitivity to weather-driven fire danger in fuel-abundant systems makes it a useful indicator for analysing relative changes in fire-conducive weather, while acknowledging that vegetation type and fuel availability modulate realised fire activity.

4. How much confidence can one have in results obtained as FFDI for regions as extensive and with such diverse vegetation as Australia, Brazil and the USA?

We acknowledge the reviewer’s concern and agree that using any single fire weather index across regions with diverse vegetation types requires careful interpretation. The FFDI is not a universal measure of fire danger; rather, it is most appropriate in fuel-abundant ecosystems where meteorology, rather than fuel availability, is the primary driver of fire risk. For this reason, our study does not present results for all global regions. Instead, we focus on Australia, Brazil, and the western United States — areas where the FFDI has previously been shown to perform well (e.g., Hoffmann et al., 2003 for the Amazon; Clarke and Evans, 2019 for the US; Dowdy et al., 2009 for Australia).

We emphasise that our study examines meteorological fire danger rather than fire behaviour or impacts directly. By focusing on the weather-driven component, we can draw robust comparisons across regions while recognising that local vegetation and land-use factors will modulate actual fire outcomes. This is why our findings should be understood as projecting changes in the fire-conducive climate envelope, rather than predicting fire activity in every biome equally.

We have already outlined in response to the reviewers point 2 that future work should explore multi-index or ecosystem-specific approaches (e.g., integrating FFDI with the Canadian FWI in boreal regions, as well as newer ML-based indices). In

addition, future assessments could move towards incorporating more direct, real-world measures of fire season change, such as burned area or fire intensity. However, these outcome-based metrics are inherently stochastic, strongly influenced by ignition sources and land-use dynamics, and therefore far harder to capture consistently. Robust projections of annual burned area have only just begun to emerge, and even these remain highly uncertain and not yet suitable for assessing fire season length within a robust uncertainty framework such as the one we present here.

5. (Lines 102-107) In general, it is expected that any meteorological danger index will be well correlated with fire activity; This does not demonstrate the usefulness of FFDI.

We partly disagree with the reviewer's implication that correlation with fire activity is not informative for assessing the usefulness of a fire danger index. In operational contexts, the ability of an index to reliably track fire activity is a core requirement for its adoption, as it underpins preparedness decisions, resource allocation, and public warning systems. In this sense, correlation with observed fire activity is not incidental but central to why indices such as the FFDI are used in practice.

That said, we agree that correlation alone is not only possible justification for selecting an index in a climate-change context, and our manuscript does not rely on correlation as the primary rationale for using FFDI. Instead, FFDI is chosen because it is explicitly designed to integrate the key meteorological drivers of fire danger (temperature, relative humidity, wind, and antecedent rainfall), is well suited to temperate and tropical vegetated systems where fuel is generally abundant, and has a long history of operational and attribution use. We further restrict our analysis to regions where these assumptions are valid, such as excluding none-vegetated, aka fuel limited, area and discuss the limitations of the index elsewhere.

We also discuss why alternative metrics, such as burned area or fire intensity, are less appropriate for the questions addressed in this study. While these impact-based variables are widely used (including by several co-authors on this paper), they are inherently stochastic, strongly influenced by ignitions, suppression, and other human interventions, and unevenly observed across regions, as we state in the methods:

"In such contexts, FW indices provide a robust proxy for changes in fire season length and the frequency of dangerous conditions, offering a clearer signal than more stochastic fire metrics such as burned area (Jones et al., 2022). Although a range of models exist to simulate physical fire behaviour, including burnt area and fire intensity (Hantson et al. 2016; Kelley et al. 2019; Haas et al. 202X), the stochastic nature of fire mean these approaches are often less suited to informing management about timing, duration, and seasonal change. Such models continue to exhibit substantial uncertainty (Burton & Lampe et al. 2024a; Hantson et al., 2020;), particularly when applied across regions and under future projections (Kloster and Lasslop, 2017). "

As a result, they do not provide a direct or stable measure of meteorologically driven fire risk and are more difficult to project robustly under climate change. In contrast, FFDI offers a stable, meteorologically based signal that is better suited for projecting

future fire-weather conditions and for examining shifts in seasonal preparedness and adaptation windows.

Within this broader framework, the demonstrated association between FFDI and observed fire activity is included as a confirmatory check that the index behaves consistently with real-world fire outcomes in the systems studied, rather than as a stand-alone demonstration of its scientific value. We have clarified this distinction in the revised manuscript to ensure that the role of correlation is not overstated and is clearly separated from the methodological justification for using FFDI.

- The authors need to clarify many other aspects of the methodology, namely:
 1. (Lines 214-215) “Throughout this study, we consider days above the “Very High” threshold, corresponding to days above a threshold of 24, as an indicator of fire season length”. Why? This relationship must be demonstrated or there must be a citation.

The use of the $FFDI \geq 24$ (“Very High”) threshold as an indicator of fire season length is grounded in both operational practice and prior research. In Australia, the Bureau of Meteorology and fire management agencies use FFDI thresholds to define levels of fire danger, with the “Very High” category representing conditions where fires are likely to spread rapidly and pose a substantial risk to life and property (Noble et al., 1980; Dowdy et al., 2009, 2018). Previous studies have shown that days exceeding this threshold strongly correlate with periods of high fire activity in fuel-abundant ecosystems, and thus provide a robust proxy for the effective fire season (Jolly et al., 2015; Hoffmann et al., 2003). Our expanded evaluation shows that the model ensemble performs well in reproducing the time spent in FFDI categories up to Very High relative to reanalysis-based FFDI, with higher categories occurring less frequently in both simulated and “observed” FFDI at the coarse climate-model resolution. By focusing on this threshold, we capture the meteorological window when fire behaviour is most extreme and when fire management resources are typically fully mobilised, making it an operationally relevant and scientifically defensible metric for assessing season length. This is covered in revised methods 2.1 “2.1 The Forest Fire Danger Index” and 2.4.1 “Increases in Very High fire weather” where we have added :

This threshold is widely used, operationally and in previous studies, to represent conditions under which fires are likely to spread rapidly and pose substantial risks to life, property, and ecosystems (Noble et al., 1980; Hoffmann et al., 2003; Dowdy et al., 2009, 2018; Jolly et al., 2015) requiring active fire suppression and a continual state of suppression readiness.

2. (Lines 254-266) the authors state that they computed 5 metrics, but in the following lines (267-270) they also state they computed additional metrics. So, how many metrics were computed?

The description on lines 267-270 relates to the metric defined on 254-266. Though we realise this could have been clearer. In the revised m/s, we have a new subsection for

each metric under revised section 2.4 “Metrics calculated” (not added directly here due to its length)

3. The metrics studied must be well-defined. However, it is not at all clear how the "burnable land surface with an increase in fire weather" is calculated nor what the relationship is between the increase in "fire weather" and "burnable land surface".

We did define burnable land in the original manuscript; however, to improve clarity and reproducibility, we have now made its role in the analysis explicit. We moved and expanded the description to Methods Section 2.2 (“Models & Data”), where it is now clearly stated that burnable land is used as a spatial mask applied consistently across all analyses:

“All analyses are performed on the fraction of burnable land, defined using the SAGE potential vegetation database (Bett et al., 2020). Non-burnable areas—including ice-covered regions and grid cells where more than 50% of the surface is classified as bare soil—are masked out, as illustrated by the grey mask in Figure 1. This burnable land mask is applied consistently across the evaluation of modelled FFDI against Copernicus FFDI, as well as for the baseline, historical, and all future Global Warming Levels (1.5°C, 2.0°C, and 4.0°C), and across all ensemble members.”

4. To understand the last metric (The number of days in the High FFDI category), it is essential to know the relationship between the FFDI categories and the incidence of fire, on a global scale (the same scale as the study)

To start with, we should highlight the reviewers own point 5 in the pervious comment. That being said, to address this point, we have expanded the manuscript to more explicitly demonstrate the relationship between FFDI categories and observable fire activity at the global scale. Specifically, we have added a global map of mean annual burned area alongside the FFDI category maps in Figure 1, and introduced a new analysis (Figure A1) examining the mean time spent in Low–Moderate, High, and Very High FFDI categories across bins of observed burned area. Together, these additions show a coherent and interpretable relationship between increasing burned area and increasing time spent in higher FFDI categories, supporting the physical relevance of the index beyond its region of origin.

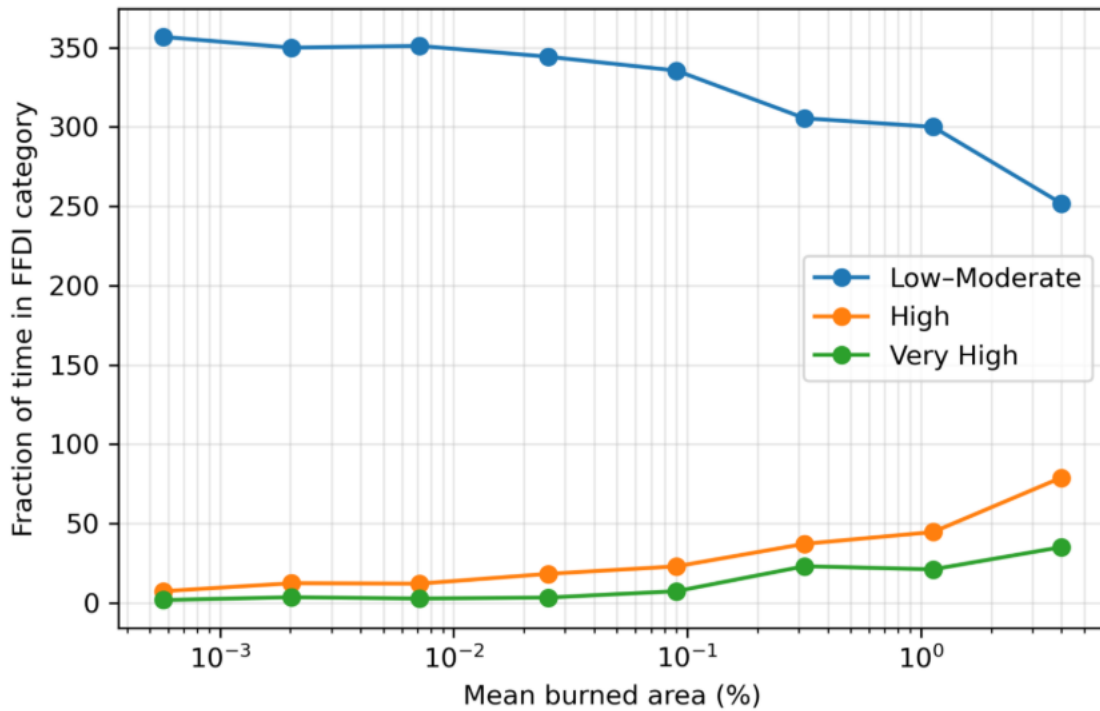


Figure A1: Relationship between burned area and time spent in fire weather danger categories. Mean fraction of time spent in each McArthur Forest Fire Danger Index (FFDI) category is shown as a function of mean burned area across grid cells. Burned area is taken from the ESA Fire CCI burned area product (version 5.1), expressed as percentage of area burned. Time spent in each FFDI category is calculated using Copernicus reanalysis meteorology. Burned area values are grouped into logarithmically spaced bins, and lines show the mean fraction of time in each category within each bin.

However, we do not consider a direct global correspondence between FFDI categories and burned area to be essential for the objectives of this study. The reviewer themselves states this their main review point 5. Burned area and other global fire activity metrics quantify the occurrence of fire but do not directly represent the human, ecological, or economic impacts that motivate concern about extreme wildfires. These impacts are strongly linked to ecosystem vulnerability, human and asset exposure, which is strongly mediated by adaptation of human and ecological systems, such as land management practices, suppression capacity, ignition sources, and socio-economic context, many of which are only weakly coupled to global fire activity measures such as burned area.

For example, regions such as the Sahel or northern Australia experience frequent, often predictable fire activity with comparatively limited impacts, while much smaller burned areas in regions where fire is infrequent, particularly those with high exposure of people, infrastructure, or vulnerable ecosystems, can result in disproportionately large consequences. As noted in recent syntheses (e.g. Hamilton et al., 2024; Jones et al., 2024), there is currently no robust global framework for mapping or quantifying these impact-relevant wildfire outcomes.

Accordingly, our use of FFDI is intended to characterize changes in meteorological fire danger and the potential for more extreme wildfire conditions, rather than to predict burned area or fire occurrence directly. The additional analyses now included demonstrate that FFDI remains physically meaningful at the global scale, while reinforcing that its primary value in this study lies in assessing changes in wildfire danger relevant to preparedness, management, and risk, rather than fire activity alone.

We have added the following note to Figure 1 to make this clear: *“Burned area is included as an observable indicator of fire activity rather than a direct measure of fire danger, and should therefore be interpreted only as contextual information rather than a target or metric of fire management effectiveness.”*

The results section also needs a lot of additional work.

5. Although the exact reproduction of the observed patterns was not expected, the evaluation of the FFDI calculated with simulations and observations (section 3.1 and Figure 1) is not convincing, for several reasons: first, only one metric is evaluated; second, even for this metric what the results demonstrate is that the simulated FFDI is similar only for Australia, the region for which FFDI was developed. The inadequacy of FFDI to assess meteorological fire danger globally is recognized by the authors, at least for boreal regions, and these results do not increase confidence in the methodology adopted and, consequently, in the results and conclusions obtained.

We agree that evaluating simulated FFDI globally requires more than a single metric and careful consideration of regional applicability. While the original comparison of Very High FFDI days already showed that the PPE ensemble captures the large-scale spatial patterns of fire weather well, we have expanded our evaluation to address the reviewer’s concern and to demonstrate the broader utility of the model.

Specifically, we have now:

- Compared the PPE ensemble to Copernicus FFDI across all four categories (Low–Moderate, High, Very High, and Severe+), providing a more complete picture of fire weather distribution and showing that the model reproduces observed patterns not only for Very High FFDI but across the full spectrum of fire danger.
- Evaluated the position of Copernicus observations within the ensemble for each category, following the method of Barbosa et al. (2025). This analysis shows that in the majority of grid cells, the observed FFDI falls within the ensemble range, demonstrating that the ensemble is well-calibrated and not systematically biased.
- Quantified model performance using Normalized Mean Error (NME) scores, which assess both temporal and spatial agreement. NME1, NME2, and NME3 indicate that the ensemble reproduces observed fire weather with substantial skill up to the Very High category, and while performance is reduced for Severe+, supporting our decision to use Very High + for proxy for fire seasons, the model still captures the broad spatial patterns. Comparisons with null models further demonstrate that the ensemble adds

substantial predictive value over simple single-value or randomly resampled baselines.

Together, these analyses provide multiple, complementary lines of evidence that the PPE ensemble is a robust and useful tool for global fire weather assessment, while acknowledging the known limitations of FFDI in regions such as boreal forests. By extending the evaluation beyond a single metric and including spatial and category-wide comparisons, we increase confidence in the methodology and in the subsequent analyses of changes in fire weather under future warming scenarios. See new evaluations, 2.3 in methods and 3.1 in results (not reproduced in responses due to length). Our new figure 1 is now:

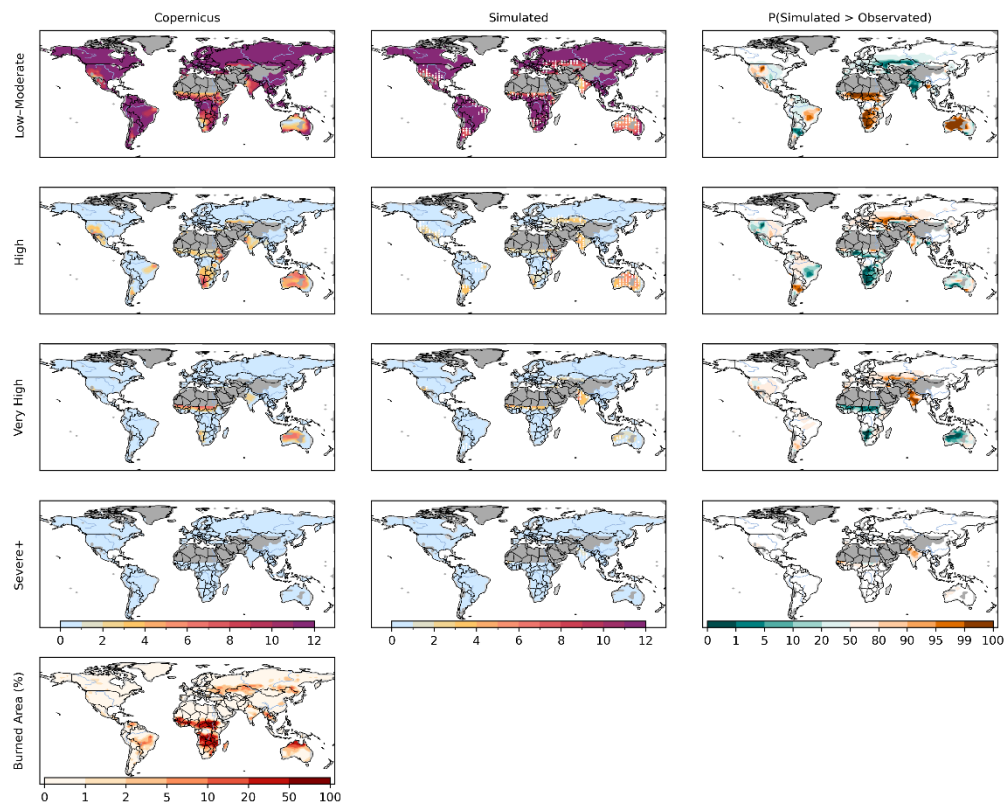


Figure 1. Comparison of the average number of days per year within the (top to bottom) Low-moderate, High, Very High, and Severe or greater FFDI categories for 1986-2005, defined in Table 1. The first column shows data from Copernicus FFDI (CMES, 2019), while the second column displays HadCM3C PPE historical data (1986-2005). The model ensemble data indicate the model agreement (consensus) across the PPE, with large white dots highlighting where more than 10% of ensemble members exhibit over one month's variation in time within that category. The third column presents the percentage of members with more time in each category compared to Copernicus, following the ensemble-based evaluation described in Barbosa et al. (2025). A value of 50% indicates that Copernicus falls in the middle of the ensemble, suggesting an unbiased estimate, 0 or 100 indicates a biased ensemble that over- or underestimates the amount of time, respectively. We show mean annual burned area from ESA Fire CCI version 5.1 [ref] in the lowest panel for qualitative comparison with global patterns of fire weather derived from FFDI. Burned area is included as an observable indicator of fire activity rather than a direct

measure of fire danger, and should therefore be interpreted only as contextual information rather than a target or metric of FFDI effectiveness.

And we have added the following evaluation table:

Table 2: Normalized Mean Error (NME) scores for the PPE ensemble compared to Copernicus-derived FFDI for the baseline period (1986–2005). Scores are shown for each FFDI category (Low–Moderate, High, Very High, Severe+), with NME₁ representing absolute differences, NME₂ differences after mean removal, and NME₃ differences after removal of both mean and variance to assess spatial patterns. Lower scores indicate better model performance. For comparison, three null models are included: the median and mean null models, which represent the best single-value predictions across the domain, and a randomly resampled null model (1000 iterations, 5–95th percentile range), which provides an expected range for a model with no skill.

	Metric	Low-Moderate		High		Very High		Severe+	
Null Model	Median	0.69		0.70		0.60		0.52	
	Mean	1.00		1.00		1.00		1.00	
		5%	95%	5%	95%	5%	95%	5%	95%
	Randomly resampled	1.08	1.16	1.1	1.14	1.05	1.11	1.00	1.04
Simulated	NME ₁	0.26	0.33	0.4	0.52	0.30	0.40	0.64	1.52
	NME ₂	0.27	0.37	0.42	0.53	0.31	0.46	0.79	2.67
	NME ₃	0.29	0.38	0.42	0.59	0.32	0.44	0.62	0.83

- About Figure 1, It is necessary to explain how the "Consensus" is objectively assessed; what do the values indicated in the legend mean? The colour scale is inadequate; it is impossible to perceive the different shades in the figure.

Because we have further expanded the evaluation for figure 1 we now no longer have any consensus plots presented in figure 1 and our approach to calculating consensus for subsequent plots is detailed in the methods section of the manuscript.

The below text responds to the original reviewer's comment but is now superseded so is not in the manuscript but we include here for completeness and clarity:

We thank the reviewer for raising this point and agree that a clearer explanation of the consensus concept and calculation is necessary. We will add in the following text in the methods to explain what the consensus calculation process is and then refer to in the figure caption

The method of model consensus is used as an approach to represent and maintain the full range of information available in the perturbed physics ensemble using the method of Kaye et al, 2012. This method calculates the level of model agreement for each grid square and thus highlights areas with higher or lower agreement which can be interpreted as higher or lower confidence in the strength of the projected change. This approach maintains more of the ensemble's spatial characteristics than taking the model mean

We use the approach from Taylor et al, 2013, based on Kaye et al, 2012 to summarise the information over the whole ensemble using a consensus mapping method. We calculated and visualised the number of days per year above the Very High FFDI threshold for each ensemble member and each grid cell. We then created 4 categories (or bins) of this number of days to classify each model response in to: 0 days, 1-20 days, 21-40 days, and >40 days. For each grid cell we count how many ensemble members have a value in each category to give the model consensus. The higher the consensus (or model agreement) the darker the colour. For figure 1 there is a high level of model consensus for the majority of grid cells, so the full range of colours are not visible on the plot.

7. Section 3.2, Figure 2. Why the results for RCP8.5 to 1.5°C appear to be less serious than for the RCP2.6 scenario to 1.5°C? This apparent discrepancy is also illustrated in Table 3.

We thank the reviewer for raising this point. One reason for the difference is that the RCP2.6 and RCP8.5 scenarios differ in the global patterns of some of the forcings, particularly aerosols. The rapid reduction in fossil fuel burning featured in RCP2.6 leads to a rapid reduction in atmospheric aerosol concentrations, and since aerosols exert a cooling effect which is more concentrated over land regions which are the source of emissions, this leads to an additional warming influence over these regions. Another reason arises from the presentation of results relative to Global Warming Levels (GWLs) rather than time periods. In our analysis, all results are shown at the point when the global mean temperature reaches a specific GWL, e.g., 1.5 °C above pre-industrial levels. For higher-emission scenarios (e.g., RCP8.5), the climate reaches 1.5 °C sooner and under slightly different transient climate conditions than in lower-emission scenarios (e.g., RCP2.6). Since regional rates of change can depend on processes with different responses dates, such as warming of the ocean vs land, or changes in vegetation and consequent feedbacks on climate, this can lead to small differences in the projected FFDI change at the same GWL, and in some cases, the increase in fire weather indices is slightly lower for RCP8.5 than for RCP2.6 at 1.5 °C.

Thus, the result does not imply that high-emission scenarios are less severe overall. Rather, it reflects differences in the forcings and transient pathway to the same global warming level, which can affect the fire weather response in the short term. This interpretation is consistent with previous studies using GWL-based projections (e.g., Dowdy et al., 2019). Indeed, this illustrates two of the novel aspects of our study (the use of more than one scenario in order to increase sample size in the uncertainty quantification, and the fact that our model includes dynamic vegetation)

8. Distributions in Figure 3 and Figure 4 are not normal. What are the consequences for your methods, results and conclusions?

We agree that the distributions shown in Figures 3 and 4 are not normally distributed. This is expected given the nature of fire weather metrics, which are bounded, skewed, and influenced by thresholds (e.g. days exceeding Very High FFDI). Importantly, none of the statistical methods used in this study assume normality.

All distributional comparisons are based on non-parametric approaches that are robust to non-normality, including ensemble percentiles (10th–50th–90th), ensemble agreement metrics, and paired one-sided Wilcoxon signed-rank tests. These methods assess changes in central tendency and direction consistently across ensemble members without relying on distributional assumptions. As noted in the Methods, we also verified that equivalent conclusions are obtained using parametric alternatives, but we retain the non-parametric tests precisely because the underlying distributions are non-normal.

As a result, the non-normality of the distributions has no negative consequence for our methods, results, or conclusions. Instead, it reinforces the appropriateness of the chosen statistical framework, which is designed to characterise ensemble spread, tail behaviour, and robustness of projected changes rather than to infer properties of an assumed Gaussian distribution.

9. Section 3.3. What is the Mitigation anomaly? Not clear.

We thank the reviewer for the comment and will adjust the name of this section and plot to make it clearer and avoid any confusion. It is an attempt to show how many less days there would be above the FFDI threshold with strong mitigation scenarios i.e. Global Warming Levels of 1.5 and 2.0 degrees, compared to a Global Warming Level of 4.0 degrees, and explore any differences between 1.5 and 2.0 degrees. We have adjusted the heading to be, 'Mitigation potential'. This is also better described in the new methods section:

2.4.2 Mitigation potential

To assess the potential benefits of limiting global warming, we calculated the mitigation potential in terms of reduction in days above the Very High FFDI threshold vs 4°C, globally and for each region. We used the same calculation for the number of days at or above Very High FFDI described in section 2.4.1. The mitigation potential was then calculated as the difference between the number of Very High FFDI days at 4.0°C under RCP8.5 and the number of days at the lower warming levels (2.0°C or 1.5°C) for the same ensemble member. This yields a distribution of reductions across the ensemble, which captures both spatial and parametric uncertainty.

Significance of the differences was assessed using a one-sided, paired Wilcoxon signed-rank test, with pairing by ensemble member. This test evaluates whether the reduction in Very High fire weather days is consistently less than zero across the ensemble for 4.0°C vs the lower GWLs.

10. The caption of Figure 4 is not clear. What is the "mitigation anomaly of the change in the annual average number of days above the FFDI threshold"? What FFDI threshold?

We have adjusted this caption figure for Figure 4 as detailed above, to be “The influence of mitigation scenarios” and include the FFDI threshold that we are referring to, which is the number of days above the Very High FFDI threshold.

11. Section 3.4. Very confusing description of the results. It is necessary to present numerical values that demonstrate what is written. Visual analysis is not enough. Are the reported differences statistically significant?

We thank the reviewer for this comment and agree that the original presentation of Section 3.4 required clarification. We have therefore substantially revised both the Methods (Section 2.4.3) and Results to more clearly explain the purpose of the random forest analysis, the numerical information provided, and how uncertainty is assessed.

The aim of the random forest analysis is not to statistically quantify or test changes in individual meteorological drivers in isolation, but rather to diagnose how the relative contributions of the constituent components of the FFDI differ between the baseline climate and a strongly warmed (+4 °C) climate in cases where fire weather itself changes substantially. In other words, the analysis addresses the question: given a change in FFDI, what combination of meteorological drivers underlies that change? rather than whether individual drivers differ significantly between climates.

To support this diagnostic interpretation, we now present numerical summaries of variable importance alongside uncertainty estimates. We employ a bootstrapping approach, generating 1000 independent random forest models per region and climate state, each trained on a random subset of days with FFDI > 12. For each variable, we report the mean and standard deviation of permutation-based importance scores across the ensemble, which are shown explicitly in the revised figures. This provides quantitative information on both the magnitude and variability of each driver’s contribution, rather than relying on visual inspection alone.

Formal statistical significance testing between baseline and future variable importance scores is performed using a unpaired one-tailed t-test on resultant variable importance scores, and shows significance of <0.001. However, strictly speaking, the significance test is not as necessary here as other analyses, as permutation-based importance values are relative measures within a given model and are not independent across predictors. Instead, the robustness of reported changes is assessed through the consistency of shifts across the bootstrapped ensemble and across regions. We have clarified this interpretation in the text to avoid implying hypothesis testing or single-driver attribution.

These revisions are intended to make clear that the random forest analysis is used as a physically interpretable diagnostic tool to understand the emergence of increasingly compound fire-weather conditions under warming, rather than as a framework for statistical inference on individual meteorological variables.

12. Line 414. Why the citation?

No longer needed because we have updated this section as detailed above

13. Line 423. Is the Loess fit described in the methodology?

No longer needed because we have updated this section as detailed above

14. Table 4. What is the meaning of the colour palette?

The colour palette is used to visually emphasize the magnitude of the values in Table 4, supporting rapid interpretation of relative changes across regions and warming levels. Such visual cues can be helpful for readers with different cognitive processing styles, including some forms of neurodivergence and specific learning differences, and are also used by members of the author team for accessibility reasons. The journal has indicated the colour may well be removed in the final version, though we keep it in the preprint enhance accessibility. We have now explicitly linked the colours to defined value ranges and added a colour bar to clarify their meaning and ensure transparency for all readers.

Table 5 Projected shift in the end-of-season date for controlled burns in Australia, Brazil, and the USA at each global warming level (GWL). Calculated as the number of days earlier in the year that the FFDI reaches the baseline threshold, typically marking the end of the burn season: 31st October for Australia, 1st September for Brazil, and 31st May for the USA. Values show the 10th and 90th percentiles across ensemble members, indicating the range of projected changes. Likelihood is based on the % of ensemble members that show an early end of burn season.

RCP2.6		Australia			Brazil			USA		
		10%	90%	Likelihood	10%	90%	Likelihood	10%	90%	Likelihood
	Historic	0	35	89	7	25	96	-7	30	72
	1.5°	-4	36	88	9	26	99	-6	28	70
	2°	10	42	93	9	51	100	-22	70	68
RCP8.5	Historic	-3	32	78	0	26	88	-3	36	79
	1.5°	5	30	92	11	27	96	-3	38	83
	2°	7	36	96	12	32	100	5	36	92
	4°	5	44	96	33	59	100	21	48	96

Later or no change	Less than a week earlier	1-2weeks earlier	2 weeks to one month	More than a month earlier
--------------------	--------------------------	------------------	----------------------	---------------------------

--

In the Discussion, the authors interpret the results but do not validate them. Comparison with the findings of other similar studies is fundamental to increase confidence in the findings of this study. Authors have to focus on what distinguishes and adds value to their study and its results. How much confidence can the reader have in the results/conclusions of this study? Why should we have more confidence in their results than in the results of other studies of this type?

We agree that situating our results within the context of existing literature is essential for building confidence in our conclusions. In the revised Discussion, we have substantially expanded comparisons with previous global and regional studies of fire weather and fire risk under climate change. These comparisons demonstrate strong qualitative and quantitative agreement with prior findings, including projected increases in fire weather severity, longer fire seasons, and sensitivity to additional warming even at low Global Warming Levels (e.g. Jones et al., 2022; Peng et al., 2023; Son et al., 2021; Abatzoglou et al., 2019).

The added value of this study lies not in contradicting earlier work, but in *strengthening confidence* in these well-established signals through complementary lines of evidence. In particular, we distinguish our approach by (i) explicitly quantifying uncertainty using a large perturbed physics ensemble, allowing assessment of robustness and confidence across emissions scenarios and warming levels, and (ii) extending analysis beyond peak fire season to the full annual cycle, including transitional and low fire weather periods directly relevant to Integrated Fire Management.

By demonstrating that key findings from earlier studies emerge consistently across a wide range of plausible climate model configurations, and by showing how these signals translate into operationally meaningful changes in seasonal fire risk, our results provide additional confidence in both the direction and relevance of projected fire weather changes. Rather than replacing existing studies, our work complements them by linking mitigation-sensitive fire weather projections, uncertainty quantification, and management-relevant seasonal diagnostics within a single, coherent framework.

- The manuscript is not well written. The title does not adequately describe the study described in the manuscript. The Introduction is very long and confusing. The text is somewhat disorganized, there is some repetition which makes the manuscript difficult and boring to read. Several parts of the manuscript must be rewritten to resolve this issue.

We have restructured the Introduction section in light of comments from both Reviewer 1 and 2, so that it is now clearer. We have kept the region-specific mitigation strategies in the Introduction, as Reviewer 1 noted they are helpful, but we have removed some details of the FFDI that were repeated in the Methods section. We have also edited the title to be clearer and more closely linked to the results of the study as follows: “Future fire weather projections show the importance of both mitigation and adaptation for dynamic fire management”

However, perhaps the most serious issue with the written style is the lack of clarity and rigour. Some examples to illustrate this criticism are as follows:

1. Fire and wildfire are not the same. Wildfires are not “defined as unusual or extraordinary free-burning vegetation events” as claimed by the authors in line 30. Fire is a wanted and controlled biomass combustion that can be used as a tool, while a wildfire is an unwanted, unauthorized and uncontrolled biomass combustion. The scientific community knows the difference and it is time to start using the correct concepts, like wildfire regime, wildfire severity, and so on.

Throughout the manuscript, we adopt the definition of wildfire used in the UNEP report *Spreading Like Wildfires*, which is cited in the first line of the Introduction. That report, developed by 52 experts from 39 organisations across 17 countries, including fire scientists and practitioners, explicitly defines, in its first line, that ‘... “wildfire” is defined as “an unusual or extraordinary free-burning vegetation fire...”’.

This definition has also been adopted in major community-led initiatives, including the FLARE project (Hamilton et al., 2024), which involved more than 30 co-authors and focused on identifying emerging wildfire challenges, and the State of Wildfires initiative (Jones et al., 2024; Kelley et al., 2025), a global synthesis involving over 60 co-authors across fire science disciplines. The UNEP definition is therefore not idiosyncratic, but reflects an established and widely used community consensus, and has been directly cited or used in numerous recent fire studies.

Throughout the original manuscript, we consistently referred to fire regimes to emphasise that regime characteristics are the primary targets for fire management and risk reduction. We have revised the text, including the title, to explicitly include the term wildfire where appropriate, consistent with the UNEP definition and to improve clarity for a broader readership. This usage does not conflate controlled fire with unwanted fire, nor does it conflict with established concepts such as fire regime or fire severity, but instead aligns the manuscript with internationally adopted terminology in contemporary fire science and policy contexts.

Hamilton, D.S., Kelley, D.I., Perron, M.M., Llorc, J., Burton, C., Liguori-Bills, N., Barkley, A.E., Buchholz, R., Diez, S., Dintwe, K. and Forkel, M., and others 2024. *The Fire science Learning AcROSS the Earth System (FLARE) Working Group FLARE: Fire science Learning AcROSS the Earth System Igniting Progress: Outcomes from the FLARE workshop and 3 challenges for the future of transdisciplinary Fire Science* (Doctoral dissertation, Zenodo).

Jones MW, Kelley DI, Burton CA, Di Giuseppe F, Barbosa ML, Brambleby E, Hartley AJ, Lombardi A, Mataveli G, McNorton JR, Spuler FR and others. State of wildfires 2023–2024. *Earth System Science Data*. 2024 Aug 14;16(8):3601-85.

Kelley DI, Burton C, Di Giuseppe F, Jones MW, Barbosa ML, Brambleby E, McNorton JR, Liu Z, Bradley AS, Blackford K, Burke E. and others State of Wildfires 2024–2025. *Earth System Science Data*. 2025 Oct 16;17(10):5377-488.

United Nations Environment Programme (2022). ***Spreading like Wildfire – The Rising Threat of Extraordinary Landscape Fires***. A UNEP Rapid Response Assessment. Nairobi.

2. Fire danger and fire risk are also not the same, but the authors seem to use these concepts as synonymous. See, for example, the abstract and the sentences in lines 38-39. The FFDI (McArthur Forest Fire Danger Index) is a fire danger or fire risk index?

We agree with the reviewer that fire danger and fire risk are distinct concepts and should not be used interchangeably. We have revised the manuscript throughout to ensure consistent terminology, using fire danger when referring specifically to the McArthur Forest Fire Danger Index (FFDI) and other fire weather or danger indices, and fire risk when referring to broader, large-scale wildfire risk that includes potential impacts.

To avoid ambiguity, we have also added a clarifying statement in the Introduction explaining that fire danger represents the meteorological and fuel-driven hazard component of fire risk, while fire risk more broadly incorporates fire danger together with exposure and vulnerability (i.e. the potential for damage to ecosystems, infrastructure, and society). Throughout the revised manuscript, FFDI is therefore treated explicitly as a fire danger index, and projections based on FFDI are discussed in terms of changes in fire danger rather than fire risk per se.

These changes improve conceptual clarity and ensure alignment with established fire science terminology.

3. (lines 85-87) the long-term climate conditions that a fire needs to burn if ignited with available fuel are not “fire weather” but “fire climate”

Changed to “Fire weather refers to the short-term atmospheric conditions, such as temperature, humidity, drought and wind that influence fire ignition and spread, and which are themselves shaped by longer-term climate conditions.”

4. (Lines 93-94) It is not the “fire indices” that “can forecast changes in the number of fire weather days and season lengths,…”

Replaced “can forecast” with “are a useful proxy for changes in the number of fire weather days and season lengths..”

5. In lines 15-16, the authors state that they “assess uncertainty in fire weather projections globally and for three regions: Australia, Brazil, and the USA”. This adequately describes the study described in the manuscript?

We have clarified the text in the abstract as follows:

To address this, we explore future climate-driven fire-weather projections using the McArthur Forest Fire Danger Index (FFDI) and a large perturbed-physics ensemble, enabling a systematic assessment of uncertainty and confidence in projected changes globally and across three focus regions: Australia, Brazil, and the United States of America. We evaluate future fire weather across all phases of the annual cycle relative to both a historical baseline (1986–2005) and a recent reference period (2004–2023), under three Global Warming Levels (1.5 °C, 2.0 °C, and 4.0 °C) and two emissions scenarios (RCP2.6 and RCP8.5). In addition to changes in season length and peak FFDI, we quantify transitions between meteorological fire danger periods and

shifts in low fire weather windows, linking projected climate change impacts directly to the timing and feasibility of Integrated Fire Management (IFM) activities.

6. In line 18, the authors state that they project the fire weather severity. How was this severity assessed?

We have removed “and severity”.

7. The authors repeatedly use slang, such as “fire weather days” when they should write “the number of days with...”

We have added to our methods: “We start with evaluating FFDI from model simulations against FFDI constructed from observations. Figure 1 shows the number of days above the Very High FFDI threshold (**which we define** as “fire weather days”)”

8. (Lines 180-181) Figure 1 does not show the indicated correlation. Show the average number of days per year above the Very High FFDI threshold in 1986-2005 for HadCM3C PPE historically and with the Copernicus FFDI for the same metric and period.

The metrics shown are consistent, and we have now clarified explicitly in the figure caption that the HadCM3C PPE and Copernicus FFDI are compared over the same baseline period (1986–2005). We note, however, that HadCM3C is a model ensemble, whereas Copernicus represents a single, observation-constrained realization. Comparing an ensemble to a single spatial field is therefore fundamentally different from comparing two single-valued datasets.

In the original manuscript, we compared Copernicus FFDI to the ensemble consensus of the same metric, which is an appropriate first-order approach for assessing whether the observed field lies within the simulated ensemble envelope. We agree that this evaluation was not substantial enough (see response to comment xxx). We have expanded the evaluation substantially. Figure 1 now includes additional FFDI categories rather than only Very High, and we have added explicit ensemble–observation comparison metrics that are designed for ensemble evaluation, including the position of Copernicus within the ensemble distribution and Normalised Mean Error diagnostics.

These additions, as well as the original comparison, follow established ensemble evaluation methodologies (e.g. Barbosa et al., 2025; Kelley et al., 2019; Burton et al., 2025). Together, these changes demonstrate that the observed patterns are generally contained within the ensemble spread and that the ensemble shows skill relative to appropriate null models, thereby strengthening confidence in the methodology and subsequent analyses.

See response to reviewers main comment number 5.

Technical corrections

The manuscript requires a large number of corrections, as a result of an excessive number of typos, errors in writing units (e.g., unit “stuck” to the numerical value, e.g., “1.5m” and “10m” in lines 199-200, but also other errors such as “km/hr1” in line 200) and even citations, for example, “by (Noble et al., 1980)” (line 102), “by (Noble et al., 1980)” (line 116), etc. Please use an n-dash instead of a minus sign to define a period.

We have gone through the m/s to check for these.

Authors should review the entire manuscript to identify and correct any existing errors.