

General response to referee comments

We thank the editor and the two referees for detailed and encouraging feedback on our manuscript that will help improve it considerably. In particular, the comments have been helpful for assessing uncertainties of our study results and for advancing clarity and rigor of our manuscript.

The main issues raised by the referees concern the interpretation of the co-occurrence measure and the selection of GPP datasets. Therefore, we will expand the description of the co-occurrence rate and discuss dataset-related sensitivities and limitations more thoroughly. Moreover, we conducted multiple additional sensitivity tests, including testing the FLUXCOM-X and FLUXCOM (RS) products in comparison to FLUXCOM (RS+METEO) that was used in our initial manuscript. Results and conclusions remain robust and similar for all three dataset choices. Irrespective of the dataset, the tests highlight higher statistical uncertainty with a shorter time frame as FLUXCOM-X and FLUXCOM (RS) start in 2001 instead of 1979 like FLUXCOM (RS+METEO). This results in a reduced emergence of significant spatial patterns with respect to the nullmodel, which is expected. As such, and considering the robustness of the results and conclusions across the different FLUXCOM products, we will continue to use FLUXCOM (RS+METEO) as the primary comparison dataset in our manuscript. At the same time, this observation is a strong argument for extending FLUXCOM-X (and FLUXNET operations) into the future to profit from benefits of an extended aggregation period for assessing climate extremes. Results and discussion of sensitivity tests will be referenced in the revised manuscript (in particular in the methods and discussion sections) and presented in full in the revised supplementary material. We will further revise the manuscript to emphasize limitations of the FLUXCOM datasets more and that they cannot be considered a “ground truth” for model benchmarking. The revised manuscript will stress that both observation-based and model datasets have notable weaknesses and better data on both sides is required. Based on the feedback from the referees, we will also report additional statistics in the text, provide requested explanations, and clarify wording ambiguities.

Below, we respond to the referee comments point by point. Original comments appear in *italic* font; proposed changes and additions to our manuscript are highlighted in **color**. Line numbers refer to the original manuscript.

Response to referee #2

This study investigates the impacts of compound temperature–moisture extremes on vegetation productivity using Earth system models and gridded GPP products. The topic is relevant, and the attempt to compare model-based and observation-based estimates of the coupling between climate extremes and GPP extremes is valuable. The manuscript suggests that compound extremes can have both beneficial and detrimental effects on GPP, but that Earth system models may overestimate the strength of the coupling compared with FLUXCOM-based estimates.

However, I have several concerns regarding the robustness and interpretation of the results. In particular, I think the manuscript would benefit from a more careful discussion of the limitations of FLUXCOM as a validation dataset for GPP extremes, a clearer justification and interpretation of the co-occurrence rate metric, and substantial improvements to the clarity of the result and discussion section. I outline these points below.

We appreciate the referee underlining the value of our study and their feedback on dataset sensitivities and interpretation, that we are taking up as described in the following.

Major comments

1. Reliability of FLUXCOM for validating GPP extremes

A central part of the manuscript relies on FLUXCOM GPP as an observational benchmark for evaluating the co-occurrence between compound climate extremes and GPP extremes. However, FLUXCOM is a machine-learning-based upscaling product trained on eddy-covariance tower observations and remote-sensing/climate predictors. It is therefore not a direct observational estimate of GPP, especially, it largely underestimate the interannual variability and may not fully capture extreme GPP anomalies (Jung et al. 2020; Nelson et al. 2024). This could strongly affect the inferred frequency, magnitude, and timing of GPP extremes. Therefore, the manuscript should more explicitly discuss whether FLUXCOM is suitable for validating GPP extremes.

We agree with the referee that FLUXCOM has important limitations that need to be considered when interpreting results and that our manuscript's mention of shortcomings in introduction and discussion (e.g., underestimated interannual variability) can be extended to give these more weight.

Unfortunately, no better alternative dataset that we know of exists for comparison. In initial exploratory tests we also included TRENDY, for which the magnitude of (univariate) extreme events was found to be unrealistically low in comparison (see response to review #1 as well). As a result, FLUXCOM was the only available gridded and observation-based GPP product that covered substantial parts of the historical period and provided reasonable extreme event distributions. To compensate the low seasonality in FLUXCOM, we employ relative monthly extreme quantiles instead of an absolute and/or year-round definition of extremes. Of course, other biases in FLUXCOM remain, especially in the tropics. Therefore, we try to avoid framing FLUXCOM as “ground truth”, but rather a different and complementary perspective. We will further strengthen that framing upon revision. In this context, our study underlines why reliable GPP reference products are essential for model benchmarking and why further improvements to FLUXCOM (or other datasets) would be highly valuable. We will make this point more explicit in the revised discussion.

Mainly, we will

- ensure consistent framing of FLUXCOM as a complementary perspective rather than a reference dataset throughout the manuscript,
- update the description of FLUXCOM in the Data section (Sect. 2), highlighting the criteria for choosing FLUXCOM and discussing potential weaknesses, and
- update the Discussion (Sect. 5) and extend the discussion of limitations of the FLUXCOM dataset and their influence on the interpretation of our results (Sect. 5.1.2, previously Sect. 5.2)

I also recommend avoiding the term “observations” when referring to FLUXCOM, unless it is carefully qualified. Terms such as “FLUXCOM GPP estimates” would be more accurate.

We thank the referee for raising this aspect. We agree, this should not be mixed up. We introduced FLUXCOM as “observation-based” but simplified this to “observations” for the sake of readability. We will use “observation-based” and “FLUXCOM GPP estimates” throughout the updated manuscript.

The authors may also consider using FLUXCOM-X at daily scale, which would better match the daily ERA5 climate data used in the study. FLUXCOM-X also includes more flux sites and has been reported to show higher consistency with independent atmospheric carbon-cycle constraints than previous

FLUXCOM versions. They should discuss how the choice of FLUXCOM product may influence the conclusions.

Thank you for these suggestions, which are in line with comments by referee #1. The temporal resolution of our study is constrained by the model output (see response to General comment 2 by referee #1), not the resolution of FLUXCOM (in principle, not just FLUXCOM-X but also FLUXCOM is available at daily resolution). We will clarify this limitation in the methods (see response to referee #1).

We have thoroughly tested and compared the effects of using different FLUXCOM products (FLUXCOM-X, RS, RS+METEO) on our results. This sensitivity assessment confirms the robustness of our results, and demonstrates the advantage of selecting FLUXCOM-RS+METEO to improve statistical robustness (as it allows us to use data from 1979-2014 in our analysis, not just 2001-2014 as would be the case with FLUXCOM-X and FLUXCOM-RS). We will revise the manuscript to address this in the data (Sect. 2) and methods sections (in particular, Sect. 5.1.2). Furthermore, we will add a new Sect. S4 to the Supplementary Information, which compares the results from different FLUXCOM products and discusses the effects of shortening the time period available for our analysis (see also reply to general comment #2 from referee 1).

2. Interpretation and reliability of the co-occurrence rate metric

The co-occurrence rate metric is central to the manuscript, but its interpretation as a response-time indicator is not fully convincing. In particular, the peak timing of the co-occurrence rate may not necessarily represent the time required for GPP to respond to a compound climate extreme.

For example, consider an idealized case in which GPP extremes occur at months $t = 2-8$ and 10, while compound climate extremes occur at $t = 1$ and 3-9. In this case, the co-occurrence rate could peak at a lag of one month. However, this does not necessarily mean that the dominant GPP response time is one month. For example, GPP extremes from $t = 3-8$ can be induced by concurrent climate extremes, while only two cases, $t = 2$ and $t = 10$, are consistent with a one-month delayed response. The metric could therefore identify a lagged peak even though the dominant relationship is concurrent.

This issue becomes problematic because the same GPP extreme event can be counted multiple times across different time lags. For example, if a GPP extreme occurs at time t and compound climate extremes occur both at t and $t - 1$, the same GPP event contributes to both the concurrent and lagged co-occurrence rates. This makes it difficult to interpret the lagged co-occurrence rate as evidence that climate extremes at $t - 1$ increased the likelihood of a GPP extreme at t .

Therefore, I recommend that the authors reconsider the interpretation of the “peak time” metric. At minimum, the manuscript should clearly state that the peak timing of the co-occurrence rate does not necessarily represent a physiological response time. The authors could also test an alternative event-matching framework that avoids double-counting GPP extremes across different lags, for example by assigning each GPP extreme to the nearest or strongest associated compound climate extreme within a defined time window.

This concern is especially relevant for Figure 2d, f, and h, where some peak timings exceed 8 months. Such long delays are difficult to interpret as direct vegetation response times, although they may partly reflect seasonal memory or legacy effects, but these effects are normally not more substantial than the concurrent effects unless there is a tree mortality (see Fig. S3 in Yu et al., 2025). The manuscript should discuss this more carefully.

Here, the referee raises an important potential sensitivity of the co-occurrence rate metric. In principle, we agree that spurious cases like the one described could affect the diagnosed peak time. However, spurious long-memory responses to individual events (e.g., to an earlier compound extreme which is in

fact a response to a later event) would average out, because we compute peak height and time after averaging all responses over the whole study period. It is extremely unlikely that we would find such a high number of events with the same long lag (e.g., peak timings of 8 months) without any mechanistic process chain linking the climate to the vegetation extremes. Similarly, a short lag does not necessarily mean that there is a mechanistic link for those individual compound and GPP events. However, by averaging over all diagnosed timings, those that have a mechanistic link should emerge.

We also tested the null model and preserved longer time segments when reshuffling the binary event timeseries for null model sampling. If there were annually re-occurring clusters of climate compound extremes, the individual compound extremes would be assigned to the same vegetation extreme. This would systematically alter our results compared to shorter segment lengths, which we do not find.

We will expand on the potential sensitivity in the revised manuscript (Sect. 3.3, Sect. 5.1.2). In particular, to Sect. 5.1.2, we will add the following:

The COR metric, specifically its peak timing, does not impose constraints on causality or imply that one GPP extreme can only relate to exactly one compound climate extreme. As such, it could be sensitive to clusters of compound extremes, leading to incorrectly diagnosed coupling between several temperature-moisture extremes and the same GPP extreme. However, spurious long-memory responses to individual events, for example, a GPP extreme that occurs months after a temperature-moisture extreme without a mechanistic connection, does not affect the average over a large number of events and long study periods. By averaging over all diagnosed peak times, those with an underlying mechanistic link should emerge.

Nonetheless, the peak time of the COR only represents the most frequently occurring response time of GPP to a temperature-moisture extreme averaged over all responses within the considered time frame. This does not guarantee a mechanistic link, individual physiological responses can scatter more broadly. Still, the most frequently occurring response is a reasonable starting point for exploring potential couplings. To avoid potential confusion, we will amend our description of the peak time and rephrase ambiguous statements throughout the text to clearly state that the peak timing indicates the most frequent response time and does not directly reflect plant physiology. For example, we will update the following statements (changes highlighted in bold):

L157-159: Significant CORs reflect the strength of the statistical coupling between temperature--moisture and vegetation productivity extremes and their peak timing indicates **the most frequently occurring response time of extreme GPP fluxes** to plant-physiologically demanding and beneficial conditions (Sect. 3.4).

L204-206: This response time does not **directly reflect plant physiology and does not** necessarily imply that a GPP extreme will always or exclusively occur within that period, **but that it is the most frequent time of occurrence for a GPP extreme**.

3. Soil moisture depths are different for ERA5 and different ESMs

The manuscript states that soil moisture is chosen because it directly reflects water available to plants. Please clarify which soil moisture depth is used. ERA5 soil moisture and ESM soil moisture represent different soil depths, and the soil depth is different among different ESMs; this difference may affect the comparison of GPP sensitivity to soil moisture. The manuscript should explain how this issue is handled or discuss it as a limitation.

Indeed, the direct link to water availability is the motivation to use soil moisture next to precipitation and the choice is motivated by a substantial amount of literature (see introduction). ERA5's soil model HTESSEL reports four layers of soil moisture and we aggregate all four of them (0-289 cm depth), because the soil moisture variables reported by CESM2 and MPI-ESM aggregate soil moisture up to depths of 850 cm and 983 cm, respectively. We agree with the referee that comparing absolute soil moisture/volumetric soil water content between the datasets would be a substantial limitation, not just because of different depths being included in the reported values but also because of structural offsets

in local long-term averages between the datasets. However, our definition of soil moisture extremes is relative and based on the distribution of each individual dataset. Moreover, monthly to inter-annual soil moisture extremes are governed by the surface soil (atmospheric drought) and root zone (plant stress) which are considered in all datasets.

We will add the soil moisture depths and our reasoning to the data section (Sect. 2) of the revised manuscript in the following way:

ERA5-provided soil water data comprises the top 289 cm of soil (Balsamo et al., 2015), whereas CLM5.0 considers a spatially varying depth of up to 850 cm (Lawrence et al., 2019) and JSBACH one up to 983 cm (Reick et al., 2013). This would strongly limit comparing absolute soil moisture between the datasets. However, structural offsets between the datasets do not affect the diagnosed soil moisture extremes as our definition of soil moisture extremes is relative and based on the individual distribution of each dataset (Sect. 3). In addition, daily to inter-annual soil moisture extremes are governed by the surface soil (atmospheric drought) and root zone (sustained plant stress), which are considered in all datasets.

We will also recommend consolidating soil moisture reporting conventions in the discussion by adding the following to Sect. 5.1.2:

Notably, however, the soil hydrology schemes of ERA5, CESM2, and MPI-ESM all report soil moisture up to different depths (Sect. 2). This is not expected to affect our relative definition of soil moisture extremes (Sect. 3), and all soil schemes include surface and root zone that govern plant productivity extremes (Schenk and Jackson 2002, Jackson et al. 1996). Because the deep soil (larger than 300 cm depth) may still constitute a relevant water reservoir in some settings (Schenk and Jackson 2002), more finely reported soil moisture layers by the ESMs would allow excluding this potential sensitivity of our aggregated analysis.

4. Improve the clarity and support of the Discussion

The discussion section needs substantial clarification. Several statements introduce assumptions about mechanisms, model biases, soil moisture memory, or vegetation–climate coupling without enough explanation or supporting references. The discussion would be stronger if the authors more clearly separated:

- what is directly shown by the results,*
- what is inferred mechanistically and supported by reference, and*
- what remains speculative.*

For example, the manuscript states that climate extremes pre-condition vegetation towards productivity extremes more strongly in ESMs than in the observational data, but the supporting evidence in the Results section should be shown and cited more explicitly.

The discussion should also include more references to support mechanistic interpretations, e.g. regarding soil-moisture memory, effective integration time, precipitation versus soil moisture as indicators of water availability. In several places, the current discussion raises interesting points but does not sufficiently explain how they connect to the results.

We appreciate this feedback by the referee, providing helpful pointers to how we can improve the discussion. In response, we will include more cross-references to results sections and relevant figures to underline which statements directly draw on our analysis, as well as further literature references. Furthermore, we will ensure throughout the discussion that we consistently and explicitly highlight hypotheses by using careful statements like “this could”, “we hypothesize”, “this suggests”.

Other comments

Ln 28: “diminishing” => “diminished”.

Ln 29: “high GPP” may be better phrased as “positive GPP anomaly” or “positive GPP extreme”

We thank the referee for both comments, we will update the manuscript accordingly.

Ln 39: The phrase “increased likelihoods of compound events” is unclear. Does this refer to an increasing trend in all compound-event types? Please clarify.

We agree, this is unclear in our statement. We will update it to read “changing likelihoods of compound extremes” instead.

Ln 47: The statement that “ESMs may continue to underestimate the likelihoods of temperature–moisture events and productivity extremes individually” needs more explanation and supporting references. It is not clear whether this refers to biases in temperature extremes, moisture extremes, compound-event frequency, GPP extremes, or their joint occurrence.

We thank the referee for the feedback. To clarify this statement and connect it better to the rest of the paragraph, we will rephrase it and the following sentences to read (changes in bold):

Assessing both on a global scale is crucial for model improvement and for developing statistical bias corrections, if necessary, to feed into climate change impact assessments. **Such an assessment relies on the predictive capabilities of ESMs for temperature–moisture and plant productivity extremes. Addressing weaknesses in their representation in ESMs may require complex and individual refinement of different process models (e.g. Lawrence et al., 2019). However, even without such refinement, relative changes in the coupling between temperature-moisture and productivity extremes can already indicate** shifting likelihoods of vegetation extremes under (hydro-)climate change.

Ln 56: The connection between plant functional types, soil heterogeneity, and root-zone nutrient dynamics and the main research question is not clear. Please explain how these structural model simplifications are expected to influence the simulated coupling between compound climate extremes and GPP extremes. This paragraph could potentially be merged with the following paragraph to improve flow.

We thank the referee for this suggestion. We will streamline this part of the introduction and combine the two paragraphs. We will also clarify that all simplifications in the land surface models contribute to a high level of abstraction in the representation of vegetation and potentially mask stabilizing or exaggerate destabilizing physiological dynamics of plants under extreme conditions by rephrasing the specific statement to read:

This abstraction of real world complexity employed by land and vegetation models may mask or exaggerate dynamics in soil hydrology, vegetation responses and carbon fluxes under extreme conditions.

Ln 65: Please clarify why this is challenging. Is the difficulty related to hydrological variability, plant physiological responses, soil moisture memory, model structure?

We will clarify this to read:

Modeling the highly variable land carbon sink is challenging because the tight linkages between water and carbon cycle require accurate process representation of (hydro-)climate variability which affects moisture availability and, therefore, carbon uptake through vegetation (Gentine et al., 2019).

Ln 107: The sentence “These differences between the ESMs are attractive for assessing the link between temperature–moisture and GPP extremes...” is unclear. Why are these differences “attractive”? What specific mechanistic insights are expected? Biases and commonalities in what?

Indeed, this statement could be more specific. The two ESMs cover a spectrum of process models and calibrations, which can provide mechanistic insights. We will reword the statement as:

These differences between the ESMs are useful for assessing the link between temperature-moisture and GPP extremes as a spectrum of model representations and calibrations can provide mechanistic insights into potential biases and commonalities.

Figure 4d, f: The different peak timings among compound-event types are interesting and deserve more discussion. What mechanisms could explain why different temperature–moisture combinations show different lag structures?

We appreciate the reviewer’s interest. In general, buffering capacities of vegetation and soils, and the growth-limiting regime (nutrients, temperature, moisture) influence the lag structure (Sect. 4.2.2). We will refine our explanation on long response times and will add an example to clarify:

Such long response times can occur when extremes produce limitations in moisture supply after prolonged drawing on stored moisture or due to preconditioning outside of the dominant growing season, or when they contribute moisture input for a future growing season. For example, a cold-dry event can be detrimental in the short term (leading to short peak timings for cold-dry-low) and high GPP extremes are only possible after recovery of plant-available moisture (leading to long peak timings for cold-dry-high), especially in mid- to high latitudes.

Figure 5: Consider adding direct labels above each subplot, such as “hot–dry–high GPP” or “hot–dry–low GPP”, to make the figure easier to interpret.

We appreciate the referee’s suggestion. We had initially tried out text labels above each subpanel, but with 16 panels, that resulted in a very text-heavy figure. Instead, we use hopefully intuitive symbols. However, we realized the legend could be further enlarged to make them more obvious to the reader’s eye. The revised figure will include a larger legend for that purpose.

Ln 367: Please explain the concept of “effective integration time” and provide a reference.

Ln 368: The sentence is confusing. It is unclear what “it” and “these” refer to. This part may fit better in the Introduction, where the authors could explain why results may differ depending on whether precipitation or soil moisture is used as the indicator of moisture availability.

We refer to the integration timescale of extreme hydrological conditions: precipitation extremes are an indicator for extreme atmospheric conditions but have little temporal persistence. Soil moisture, on the other hand, integrates sustained atmospheric conditions and plant stress. We will add the following explanation to the methods section (Sect. 3.1) instead of the introduction to be able to refer to all relevant concepts:

It [precipitation] influences (extreme) atmospheric hydrological conditions, which have higher short-term variability and less long-term persistence and thus a shorter memory than soil moisture. As such, precipitation represents the fast changes in atmospheric dynamics. Soil moisture, on the other hand, accumulates, that is integrates, the precipitation signal over longer timescales and is exposed to additional influences like evaporation during that time.

Further, we will adjust these statements (previously L367 and L368) to include a brief explanation, a cross-reference to Sect. 3.1 and update them to clarify what “it” and “these” refer to:

Precipitation abundance or absence, as an alternative indicator of moisture availability, reflects a shorter integration time compared to soil moisture (Sect. 3.1). This also lessens the indirect temperature control on wet/dry extreme conditions as temperature-driven evaporation directly affects soil moisture extremes but only indirectly precipitation extremes.

Ln 375: Please explain explicitly what is meant by “different mechanistic aspects”. Which mechanisms are being contrasted?

We are referring here to the process described in the beginning (and previous) paragraph of this section that fundamentally differentiates precipitation from soil moisture: the involved timescales/ integration times as precipitation extremes reflect relatively short-term water availability in the atmosphere that does not necessarily mean that this rain will become available to the plant. Soil moisture extremes, on the other hand, occur in the response to long-term accumulation and reflect water available to the plant in the soil that is more likely to persist even under hot conditions, thus supporting further plant growth. We will expand on this statement (by expanding on the example of hot–wet extremes from the paragraph) to make this clearer:

The mirrored patterns disappear for temperature-precipitation extremes, because soil moisture and precipitation emphasize different mechanistic aspects, that is, short-term precipitation extremes do not necessarily translate into water (un)availability for plant growth in the same way that soil moisture does. For example, for a hot-wet extreme defined by precipitation, hot conditions can cause rainwater to re-evaporate rather than infiltrate into the soil where it would become available to plants. In contrast, a wet extreme defined based on soil moisture indicates a high amount of water available to the plant in the soil.

Ln 380: The manuscript should further discuss why maps based on temperature–precipitation extremes exhibit strong dipole patterns.

We agree with the referee that this merits further discussion. We will expand the description of these dipole patterns in Sect. 4.4 to include brief explanations for why these differences between temperature–precipitation and temperature–soil moisture extremes exist. Furthermore, we will add to the discussion Sect. 5.2 on contrasting patterns between precipitation and soil moisture, where we omitted the “dipole” patterns previously.

Ln 409: The manuscript argues that models and FLUXCOM agree at the regional scale. Please provide stronger evidence for this claim. For example, the authors could include a supplementary version of Figure 4C–F separated by IPCC regions.

In Fig. 4c–f, each data point already represents one IPCC region. As Fig. 4 shows and is discussed throughout Sect. 4.2, for most of the regions, models and FLUXCOM agree, with some regional spread. This spread is larger for some event type combinations (specifics are mentioned in Sect. 4.2), as we will highlight in the revised sentence:

Models and FLUXCOM agree most at the global and regional scale, despite model-specific biases and spread for some event combinations (Sect. 4.2, Fig. 3, 4).

Ln 422–425: The claim that “climate extremes pre-condition vegetation towards productivity extremes more strongly in the ESMs than in the observational data” needs clearer support from the Results section. Please also clarify the logical connection between the first two sentences of this paragraph, as the current writing is difficult to follow.

We will add a reference to the results section which describe the differences between ESMs and observation-based data and that the diagnosed coupling between climate and vegetation extremes tends to be stronger in the ESMs compared to FLUXCOM. We will also improve the flow of the paragraph, such that the beginning of the paragraph will read:

ESMs mostly underestimate the intensity of daily precipitation extremes and underestimate the dry day frequency in comparison to observations (Seneviratne et al., 2021; Norris et al., 2021). In this context, it may appear counter-intuitive that locally the statistical coupling between temperature–moisture extremes and vegetation productivity extremes is stronger in the ESMs than in the observation-based data (Sect. 4.2.1). However, this may well be due to deficiencies in the land models, particularly the soil representation, rather than being related to underestimated high and low precipitation extremes.

Ln 426–427: Please clarify what type of biases are being discussed.

Here, we still refer to the same biases as in the whole paragraph so far of models tending towards stronger coupling between climate and vegetation extremes than FLUXCOM. To make this explicit, we will rephrase this to:

That we find these biases of models overestimating the coupling between temperature–moisture and vegetation extremes for extremely dry conditions (Sect. 4.2.1, Fig. 4) could point to oversensitive evaporative feedbacks, exaggerated drainage, or too low soil water holding capacity, leading to accelerated evaporation of soil moisture.