

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. 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 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 #1

This manuscript evaluates the co-occurrence of temperature-moisture extremes and gross primary productivity (GPP) extremes simulated by Earth system models (ESMs) against the FLUXCOM product. The authors calculate co-occurrence rate using a binary approach, combining with a significant test. They present an overestimated coupling between temperature-moisture and GPP extremes, especially in low latitudes. I particularly like the binary method and the significant test presented in this manuscript. The binary solution is simple but can effectively capture the link between two types of extremes, nicely bridging the modelled and observation-based results. In addition to soil moisture extremes, they also include precipitation as moisture metrics, which offers another interesting perspective of moisture propagation. In addition, I however find the manuscript needs major revision to be clearer and more informative, as outlined below:

We appreciate the referee’s encouraging feedback on scope and methodology of our manuscript, as well as their helpful feedback addressed in the following.

General comments:

1. The key metric of this paper, i.e., the co-occurrence of temperature-moisture and GPP extremes, is somewhat not clearly defined from the beginning of the manuscript. The authors use several statistical labels to indicate this co-occurrence, such as statistical links, coupling strength, etc. It would be beneficial to clearly address the key research question in the introduction and use the principal metric (currently there are statistical links/relationships, coupling strength, etc.) consistently.

Thank you, we appreciate this feedback and agree that clarity of terms used is vital and should be improved. We employ the co-occurrence rate at lag t as a metric for the statistical coupling between extreme types at lag t , and its maximum over lags 0 to 11 as a metric of the strength of this coupling (“coupling strength”). To improve the clarity of the description, we will revise the introduction to add the co-occurrence rate metric more prominently and draw the links between this metric, “statistical coupling”, and “coupling strength” early on as well. We will replace “statistical relationship”, “statistical relation”, and “statistical link” with “statistical coupling” throughout the text to reduce the number of terms that mean the same. In the updated manuscript, we will also rephrase the study description in the Introduction to make our metrics explicit.

2. One of the main findings is that the co-occurrence of both extremes is overestimated by models in low latitude regions. However, this conclusion is drawn by comparing the modelled result with FLUXCOM, which is a product using upscaling methods. Given the limited data availability, FLUXCOM GPP is considered highly uncertain over the tropical regions (Nelson et al., 2024). It is probably true that ESMs also fail to represent the co-occurrence of both extremes in low latitudes, but if there is already large uncertainty in the reference dataset, this conclusion seems to be less convincing. Although there is no directly observed GPP dataset, maybe showing complementary analysis by using another independent GPP product will show robustness of the main results. Additionally, I wonder why the authors choose to use FLUXCOM GPP instead of FLUXCOM-X (Nelson et al., 2024) as the reference data. FLUXCOM-X provides daily GPP that matches the temporal resolution of meteorological variables analyzed in this paper. Also, avoid calling FLUXCOM “observations” as FLUXCOM GPP is not observed.

We acknowledge the referee’s concerns about data quality in low latitudes and about observation-based GPP products. We agree that improvement to FLUXCOM GPP would be desirable especially in low latitudes, at the same time it appears very challenging to obtain reliable observation data in this region. We initially tested TRENDY products, too. However, we found the magnitude of (univariate) extreme events in TRENDY unrealistically low in comparison to both models and FLUXCOM as variability in TRENDY data was comparably small. As such, TRENDY data was unfortunately not suitable for our study. We do not know of any other observation-based, gridded, and global dataset that could work with our analysis. To address the comment, we will now also consider the additional FLUXCOM products, FLUXCOM-RS and FLUXCOM-X, in the revised manuscript. Additionally, we will discuss the limitations of FLUXCOM, especially with regards to the tropics, and how they may affect our results in far greater detail (in the Data Sect. 2 and the discussion Sect. 5.1.2 on potential biases in the datasets as well as in an additional Section S4 in the Supplementary with supporting figures).

We selected FLUXCOM instead of FLUXCOM-X, because its comparatively longer time coverage (1979-2014 vs 2001-2020, start/end dates included) provides more robust statistics for evaluating extremes. For comparing to CMIP6 cycle historical simulations, the usable time frame of FLUXCOM-X is further shortened to 2001-2014. We have explored this option and repeated our full analysis with FLUXCOM-X, FLUXCOM-RS+METEO, and FLUXCOM-RS for the period 2001-2014. The same was done for FLUXCOM-RS+METEO for a different period of the same length (1979-1992) to assess sensitivities with respect to the choice of time period. We will add supporting figures (new Fig. S6 –

S9) to the Supplementary Information that showcase these sensitivity tests analogous to Fig. 4. The tests confirm that our results on (dis-)similarities between observation-based and model datasets are robust towards the choice of FLUXCOM product (X-BASE, RS, RS+METEO). These are also not due to the particular time period as results for 2001-2014 (Fig. S6) and 1979-1992 (Fig. S9) for RS+METEO are very similar. The sensitivity tests also confirm that the longer 1979-2014 time frame yields more robust statistics because it reduces noise. The revised manuscript will contain descriptions, results, and discussion of these sensitivity tests (in updated Sect. 2, Sect. 5.1.2 and Sect. S4 of the Supplementary Information).

Regarding daily GPP, the limiting factor is that CMIP6 historical GPP fluxes are only provided at a monthly resolution. This necessitates limiting our comparative analysis to this resolution, too. We will include this explanation in the methods (Sect. 3.1) and add a suggestion on publishing daily GPP output in the discussion (Sect. 5.1.2).

Finally, we agree, FLUXCOM should not be understood as an observation. In our original manuscript version, we had explained the background of FLUXCOM in Sect. 2 and subsequently replaced “observation-based dataset” with “observations” for readability. However, we understand this can be confusing. Therefore, we will now consistently use “observation-based” and “FLUXCOM GPP estimates” throughout the manuscript in-line with a suggestion by referee #2.

3. When comparing different types of temperature-moisture and GPP extremes, I found it is challenging to put the results into perspectives. It would be helpful to 1) provide statistics. How much is the overestimation? Can you give a number, e.g., in percentage? 2) summarize key findings. Which type of temperature-moisture and GPP extremes contribute to the most uncertainties?

Thanks, this observation is helpful to improve the clarity and presentation of our results. We understand the comment primarily refers to the results on dataset biases and regional spread (Fig. 4). Because discussing all the eight extreme combinations in the text would be overwhelming, we will add two tables summarizing the statistics comprehensively and refer to them in the text. We will also expand the statistics mentioned explicitly in the text by adding the most prominent over-/underestimations (Sect. 4.2).

4. Overall scientific writing needs improvement. This is listed as a general comment because many times the structure and phrasings are misleading. For instance, the last paragraph of introduction reads more like a summary. The end of introduction usually briefly summarizes the aim and the approach, and clearly define the research gap or hypothesis. What the study found is usually not part of it, rather belonging to the results and conclusion.

Several phrasings are informal and ambiguous, for instance, “bracket” (L103), “attractive” (L107), “propagate” (L119; this is often called “upscale”); Other cases are listed in the specific comments. Please revise the manuscript throughout.

We will revise phrasing throughout the manuscript, in line with this comment as well as specific comments by both reviewers. In addition, we will update the last paragraphs of the introduction to clearly present aim and methods, as well as structure of the study.

Specific comments:

1. Terminology: be consistent - statistical relationship (L10), coupling strength (L11), statistical links (L71), statistical relation (L79), statistical coupling (L116).

We appreciate the referee’s feedback that they find the number of different terms confusing. We will replace “statistical relationship”, “statistical relation”, and “statistical link” with “statistical coupling” throughout the text to improve clarity.

2. L18-19: *This sentence seems to be disconnected from the main context of this study. Please consider revising.*

Thanks, indeed, the last sentence of the abstract can be better connected to the previous one. Upon consideration, we decided to remove it completely.

3. L44: *Do you mean “co-occurrence of temperature-moisture and vegetation extremes”?*

Indeed, thanks for spotting the missing word, we will insert it in the revised version of the manuscript.

4. L51-60: *the paragraph on LSM simplification/cascading effects is not well integrated with the study objective; tighten or better link to the research aim.*

We agree the link to the objectives should be strengthened and will change it in the revised manuscript to read:

Abilities of ESMs to simulate impacts of weather extremes on vegetation are influenced by the way their land surface models (LSMs) represent plant traits.

5. L75: *What does “mitigating effects” imply? Do you mean increased GPP after temperature-moisture extremes?*

Indeed, we refer to compound extremes that raise the probability for positive GPP anomalies and will rephrase this upon revision to be clearer. We will update the sentence to:

..., we study whether compound events increase the likelihood of subsequent GPP extremes, for both negative (damaging) and positive (beneficial) anomalies.

6. L136: *The recent released FLUXCOM-X product provides GPP at daily resolution (Nelson et al., 2024). This can resolve the mismatch of time scales between different datasets. I think the authors at least need to discuss to which extent the choice of reference data might influence the results.*

As discussed in the response to the general comments, the models are the limiting factor in the temporal resolution of the GPP data. We will make this clearer here upon revision of the manuscript. We will also include results using FLUXCOM-X in the updated manuscript for sensitivity testing and update this section accordingly. As a result, at this location, we will change the text to read:

...because ESM-provided GPP data is available only at monthly resolution, we also use monthly FLUXCOM products despite daily resolutions being available.

7. L144: *Unclear phrasing. Please clarify what “characteristic response time” and “demanding and beneficial conditions” mean.*

To improve the phrasing, we have changed the sentence to read (and refer to the specific subsection for further explanation):

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 conditions that are detrimental and beneficial for plant growth (Sect. 3.4).

8. L146: *How did you remove the seasonal trend? Please clarify the calculation processes.*

The seasonal trend is removed by defining extreme conditions based on monthly event distributions. With the extreme statistics being sensitive to many means of signal transformation, this has the advantage of keeping the monthly event distribution unchanged. It also does not affect the relative threshold/quantile-based event ranking across months of different years, which is essential for our extreme event definition. We will update the description of the methodology and add this reasoning to the methods.

We define extreme events based on monthly thresholds and local extreme conditions, which implicitly removes seasonal trends.

9. L149: *“Preliminary tests...”*: If you make the claim without directing readers to tests or supporting figures, ensure the methods or supplement documents enough detail so reviewers can judge its validity.

To diagnose the relationship between temperature-moisture and GPP extremes, we only consider individual events within the same background state (as we consider coupling of events only up to 12 months apart). Therefore, the individually diagnosed relationship is unrelated to long-term trends. Furthermore, our sensitivity test comparing an early and late period of our data show the robustness of our results in spite of the long-term trend (newly added Fig. S6 and S9).

Early on in our analysis, we nevertheless tested different detrending methods and whether they add value to our analysis. We found long-term detrending to distort GPP autocorrelation which affects the diagnosed duration of GPP extremes and, thus, the diagnosed coupling. Therefore, artifacts from long-term detrending would impair our analysis more than detrending could help our research question. We will add this explanation into the methods section (Sect. 3.1):

We deliberately keep the anthropogenic CO₂ trend in the data, because we are mainly interested in the mechanistic coupling between temperature-moisture and GPP extremes and not in potential state dependencies of this coupling. As we only consider events within the same background state, individually diagnosed coupling is unrelated to long-term trends. Furthermore, a sensitivity test comparing an early (1979-1992) and late period (2001-2014) of our data shows the robustness of our results despite the long-term trend (Fig. S6 and S9). Preliminary tests further indicated that artifacts from detrending outweigh the benefits of reducing non-stationarity, as it distorts the autocorrelation of GPP time series.

10. Figure 1: *This could be a nice figure to understand the method. Currently, the fonts are very small and affects the readability. Panel (c) is not clear to illustrate the calculation of the co-occurrence rate. Is the overlap measured by the product of the two series above? What do the numbers next to the arrows mean? How is the time lag considered when calculating the product?*

We appreciate the referee finding the figure useful to understand the methodology. We will increase the font size as much as possible.

Referring to panel (c), the referee is correct in assuming the co-occurrence rate builds on the product of the two binary time series. Thanks to this comment we realized the numbers next to the arrows (representing that product) are superfluous, because they mirror the “overlaps”/“co-occurrence” shown below. We will remove these numbers in the revised version of the figure. As to the lagging, the computation is repeated for every lag increment between both binary event time series. We will add this information to the panel.

11. L165: *“reduces effective integration timescale”* is not clear. Similar phrase “moisture integrating time” appears again at L365. I suppose you mean precipitation show higher variability thus influences hydrological conditions with shorter memory, compared to soil moisture that is accumulated over longer-term. Please clarify.

Yes, exactly. We are revising the phrasing to also clarify the usage of integration time, taking the suggested rewording into account:

It [precipitation] influences (extreme) atmospheric hydrological conditions, which have higher short-term variability and less long-term auto-correlation 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.

We will also update the statement in L365 to read:

Precipitation abundance or absence reflects a shorter integration time of moisture availability compared to soil moisture as an alternative indicator (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.

12. L206: The comparison metric between model and observation-based data is mentioned only briefly. As this is a major part of the results, it is important to prepare the authors which kind of metrics are used to evaluate model performances. For instance, how is the correlation in Figure 3 and Figure S2 calculated? Why do you choose correlation instead other metrics, e.g., RMSE to evaluate model performances?

We agree that it will be beneficial to readers to explain the methods used for this comparison in more detail. Pattern correlations are computed by pair-wise Pearson correlation of all significant grid cells. They measure strength and direction of similarity while ignoring differences in scale or offsets between datasets. This is ideal to evaluate the similarity of shapes (“patterns”) when datasets include localized offsets/biases, as it is the case here. In contrast, RMSE or MAE quantify the absolute distance between a prediction and a ground truth and are therefore more “demanding” on the dataset(s) skill(s). If we were to diagnose high pattern correlations, looking further into RMSE/MAE would be fruitful for further skill quantification. However, our results mostly show low local pattern alignment between the datasets. Next to this reasoning, RMSE frames one reference dataset as a ground truth. However, as has been discussed, both ESMs and FLUXCOM have notable shortcomings. We want to avoid the ground truth framing for FLUXCOM and feel encouraged in doing so by the reviewer feedback on this topic.

We will add a paragraph to this section, describing the methods used for comparison and explaining the reasoning behind. The new paragraph also mentions how the pattern correlations are computed and why we employ them:

Pattern correlations are computed by pair-wise Pearson correlation of all significant grid cells. They indicate strength and direction of similarity while ignoring differences in scale or offsets between datasets. This allows evaluating the similarity of shapes (“patterns”) when datasets contain localized biases.

13. L216: keep wording consistent—use “temperature-moisture” unless specifying precipitation explicitly.

Thanks for spotting this mistake. Indeed, this was supposed to say “temperature-moisture” here. We will correct this and check for consistency throughout the manuscript.

We investigate the statistical coupling between temperature-moisture compound extremes and GPP flux extremes in CESM2, MPI-ESM1.2, and FLUXCOM observations.

14. Figure 2: Are non-vegetated areas masked out from the analysis? For example, it does make sense that there are co-occurrences between compound and GPP extremes in the Sahara Desert in MPI-ESM.

Areas are excluded if they do not produce a meaningful (that is non-zero, non-missing) GPP flux. Areas with (very) low vegetation coverage are included and reveal differences in model calibration and behavior (as is the case for the Sahara and discussed within the manuscript, e.g., in Sect. 4.3).

15. L255: specify where the increase is shown.

The increase is shown in the referenced Fig. 4a, b. There, both for hot-dry-low and cold-wet-high, an increase in co-occurrence rate is apparent from zero lag to a lag of one month. We are including this aspect in the text and an additional reference to Fig. 4a in the subsequent sentence.

As an example, Northeastern South America (NES) shows a rapidly increasing co-occurrence rate of hot–dry or cold–wet compound extremes, peaking at a lag of one month (Fig. 4a, b). In this region, which is characterized by a tropical non-seasonal climate, low moisture and high temperature conditions appear to have an up to eight-month legacy impact on vegetation productivity in MPI-ESM and CESM (Fig. 4a).

16. L272-273 & L277-278: *direct readers to specific figures that support these statements.*

We thank the referee for the feedback, here we are still referring to Fig. 4c, e, as is the case for the whole paragraph. To make this explicit, we will add additional references to the individual panels throughout.

17. L299: *Unclear statements. Please explain how seasonality can influence the response times. What does “stimulate supplies” mean?*

Thanks, indeed, this statement can be clarified. Essentially, we are describing preconditioning effects through temperature-moisture extremes. Whenever there is a dominant growing season (i.e., with high plant productivity) and moisture extremes fall outside of that season (i.e., into a period with low productivity and low variability), they might still influence GPP in the upcoming growing season. This is possible both for extremes that foster damaging or beneficial conditions for plant growth. “Stimulating supplies” refers to the second case of moisture input ahead of the growing season which subsequently promotes productivity. We will revise the statement to better reflect this reasoning and improve the phrasing.

Such long response times can for example occur when extremes produce limitations in moisture supply outside of the dominant growing season.

18. L380: *What does “dipoles” mean?*

Thanks for raising the question, we will clarify this. We refer to predominantly zonal, bimodal patterns with beneficial (damaging) effects of temperature-moisture extremes in the low latitudes and damaging (beneficial) effects in the mid- and high latitudes (e.g., Fig. 5m). We will change the statement, among others, replacing “dipole” with “bimodal pattern”, such that it reads:

Instead, they exhibit a strong bimodal pattern with beneficial (damaging) effects of temperature-moisture extremes in the low latitudes and damaging (beneficial) effects in the mid- and high latitudes.

19. L466: *be specific with weaknesses. Data quality? Limited observations?*

Drawing on Jung et al. (2020) and Nelson et al. (2024), it is both: Enlarging the number of tropical FluxNET sites is expected to increase data quality significantly and input information is the major constraint on FLUXCOM data products. Additionally, furthering machine-learning methodologies and enhancing the predictor space will likely increase data quality. The fact that we diagnose substantial spread in extremes between models, between models and FLUXCOM data, and, crucially, between models and newer FLUXCOM-X data (which includes an enhanced predictor space, Nelson et al. 2024), but cannot fully pinpoint the biases further is another strong argument for increasing data coverage in the tropics. In the revised manuscript, we will make the weaknesses explicit such that the statement now reads:

However, FLUXCOM tends to overestimate tropical mean GPP and net ecosystem exchange, partly due to a lack of FluxNet sites and available measurements in this area (Jung et al., 2020; Nelson et al., 2024).

20. Sections 5.1 and 5.2 both discuss similarities and biases; clarify why they are separate or merge.

We appreciate this suggestion. Section 5.1 focuses on similarities and differences in the results whereas Section 5.2 discusses limitations of the underlying datasets. We do agree though that they are similar in that they both discuss differences between the datasets and their reasons. As such, we have combined the sections into a new Sect. “5.1 Similarities and differences between observation-based and model data” with two subsubsections: “5.1.1 Statistical coupling between compound climate and GPP extremes” and “5.1.2 Dataset limitation”. This prevents having an extremely long section and makes the differences in the discussed topics explicit.

21. L470-474 and L508-514: these read as limitations-consider merging them into one coherent limitations paragraph.

In the revised manuscript, we will expand on the limitations of COR as a metric and rework it into a longer paragraph. For completeness sake, we will keep mentioning the possible future work (previous L508-514) in that Section, where other possible extensions of the work are also discussed.

22. Figure 6: The calculation of soil moisture rate should be included in the method section.

The maximum of the soil moisture rate of change is computed as the maximum of the first derivative of the soil moisture time series at each grid point. Because this simple computation is only relevant for this discussion figure and does not require a verbose explanation, we will add this explanation to the caption of Figure 6, which will then include:

Rates of change are computed as the first derivative of the soil moisture time series at each grid point.

Technical comments:

1. L244: Do you mean “lagged time” instead of “lead time”?

We took the perspective of the GPP response here, thus, temperature-moisture extremes lead that response. However, since we use the opposite framing (lag = negative lead) in combination with “peak timing” elsewhere in the manuscript, we are rephrasing this sentence:

For example, MPI-ESM simulations show peak timings of 6 to 12 months in northern North America, whereas CESM2 tends to produce occasionally longer peak timings across all regions (Fig. 2b, h).

2. L246-247: The two sentences are repetitive.

Thanks for spotting, we will combine them into a single sentence:

This is also reflected in low pattern correlations of the peak timing (Fig. 3b).

3. L251: “Spatially resolved” -> “gridded”

Thanks, we will change this.

4. L423-424: Unclear sentence

Thanks for this feedback. We are rephrasing the statement as follows:

In this context, it may appear counter-intuitive that the statistical coupling between climate extremes and vegetation productivity extremes is stronger in the ESMs than in the observation-based datasets (Sect. 4.2.1).