nature portfolio

Peer Review File

Changes in event soil moisture-temperature coupling can intensify very extreme heat beyond expectations

Corresponding Author: Professor Douglas Maraun

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The paper investigates the differential impacts of climate change on moderate and very extreme heat events using large ensembles. It finds that very extreme heat events, with return periods of 200 years, respond differently to climate change compared to moderate heat events, largely due to changes in soil moisture-temperature coupling during heat events. The study highlights significant regional variations and uncertainties in these changes, emphasizing the need for improved climate risk assessments and adaptation planning.

The paper would benefit from a more detailed explanation of the methodology, particularly the calculation of soil moisture-temperature coupling and return periods, to enhance clarity and reproducibility.

1. My major concern is about the method to calculate return period.
The authors combine all the ensemble realizations (30) and have one return period (30x25=750). I think each simulation should be treated independently and separately when calculating the return period. Each simulation is independent and not a continuous 750-year simulations. I suggest calculating the return periods for each simulation and analyze the mean and variance of those return periods across all simulations. For example, the variances could be used for statical significance of Figure 2. As shown in Figure7, the return value differences can be very different across ensemble members.
Please provide some literatures if the method is commonly used.

2.The method of calculating coupling strength is unclear. Specifically (1) Is it correlation between Tx and latent heat or soil moisture? Is the seasonal cycle removed? Please provide equation and examples for positive/negative coupling. How Tx-ET strength relate to soil moisture-temperature coupling for general audiences of Nature Communications (i.e., why not directly use soil moisture-tempearature). (2) For correlation across maximum temperature in each year, I think the naming "event" coupling is misleading. It seems more like cross-events relationships.

3.What's the definition of "total soil moisture content" (Figure 5 b,c)? Is it root-zone soil moisture or total soil moisture in all soil layers? Does deep soil moisture reflect to the occurrence atmospheric surface conditions, or does it reflect to the previous day-to-season weather or hydrological changes (ice, groundwater).

4.Additionally, addressing the regional uncertainties and providing more context on the implications for climate risk assessments and adaptation strategies would strengthen the overall impact and relevance of the findings.

Minor comments

1.Please ensure consistent use of return time(s) or period(s), also return value(s) or level(s).
2.Please specify the raw spatial resolution of the models and the regridded resolution used in this study.
3.Please make sure refer to figures, and tables, including supplementary information explicitly in the context.
4.Line 87 Project changes "of"-> in
5.For figure 3, please add labels for x- y- axes to the plots for clarity.
6.Line 270-272 realisations-> realizations?
7.Line 278: For the warm-season means, how about area in non-tropics?
8.Check if Txx is a typo of Tx.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Title: Changes in event soil moisture-temperature coupling can intensify very extreme heat beyond expectations

In this article, the authors use climate model simulations from large ensemble runs to assess heat extremes with large return periods of 200 years relative to heat extremes with shorter return periods of 2-years. Then, they explore the source of the differences between the events with a 2-year return interval vs. events with a 200-year return interval, finding that soil moisture-temperature coupling plays a key role.

Overall, I think this is a very interesting study, and the methods are straightforward and justifiable. I have some suggestions that I recommend the authors address prior to publication, as outlined below:

Major comments
[1] I think it would be nice to more explicitly describe what constitutes an extreme event using return periods. I question whether a 2-year event is considered "extreme", as it seems more like normal, frequent event. Is an event with a 2-year interval really a "moderate" extreme event? Furthermore, why do the authors choose to compare a 2-year event to 200-year events? Why not along a gradient of return periods? Analyzing just the end points seems like a missed opportunity.

[2] L135-145, I am a bit confused how moving from a transitional to dry regime would dampen a heatwave. Even though the coupling is strongest in transitional regimes, this does not imply that heatwaves are amplified in transitional regimes, right? If a transitional regime, for example, became water-limited, all the available energy would go to sensible heat, heating the surface and increasing near-surface air temperature more than if there was any soil moisture available. It might benefit the paper to think about the distinction between transitional regimes in space vs. time. Also, it could be interesting to include a conceptual figure using example data showing how event coupling is calculated in the SI.

[3] It is unclear what is considered the "warm season", and if seasonality during the warm season affects the coupling strength, i.e. is the seasonality driven by seasonality or daily anomalies?

[4] Figure 4, The arrows below the conceptual figure are confusing. Do the arrows correspond with the conceptual figure above? More importantly, though, I am not sure this figure adds much to the paper, as the information seems to be easily described in text.

Minor comments
[1] Introduction: I think the readability would improve if the one sentence paragraphs in the introduction were incorporated into larger paragraph.
[2] has to have
[3] L61, These seem more like the key results, not your hypothesis.
[4] L84, "of our results" should be more specific
[5] L85, check grammar
[6] The language in the last four paragraphs feels too casual and vague, "good news", "bad news", "huge model uncertainties", "But still".

(Remarks on code availability)
It appears that there is code to create each figure. There is no README file describing each file, though.

Reviewer #3

(Remarks to the Author)
This manuscript explores the different responses of extreme heat events and moderate extreme events in climate change through various SMILES, and how changes in soil moisture temperature coupling affect these extreme events. The author has reached consistent conclusions on different models, providing profound insights into the response to extremely extreme heat events. However, I believe there is room for further improvement in the content. I suggest considering this manuscript, but it requires some modifications.
L73: Warm season is not clearly defined in the text, which specific months is it?
L80: Why use these three large sample models? Have you considered other large sample models? The physical process simulations between different models are inconsistent and may result in inconsistent results.
L88: The article does not explain why SSP585 is used to continue the results of the model after 2014. Will the response of extreme heat events to soil moisture temperature coupling be inconsistent under lower emission scenarios? Furthermore, I believe that the period representing the 'present' should be chosen to be 2024 or 2023, rather than 2014 when the Historical scenario ends. The SSP scenario can be used to extend the subsequent years. And, there is no uncertainty-related

information in the graph, do all grid results have a high level of confidence? I hope the author takes into account the issue of uncertainty when analyzing.
L94: To characterize the correlation between two variables, confidence analysis should be added instead of just presenting correlation data. In addition, almost all the data in the article does not include uncertainty ranges, which has caused me great concern.
L134: What conclusions did the seasonal average coupling strength of these models reproduce from previous studies? There is no necessary explanation for Figure S3, making it difficult for readers to understand its meaning.
L140: This is very counterintuitive. Humidity should always have a restraining effect on temperature drop, as it reduces the evaporative cooling effect on the surface, leading to an increase in temperature. Although the strength of land-atmosphere coupling decreases, the temperature should be increasing. I hope the author can provide clearer explanations.
L146: As above, the uncertainty range should be added to the graph, and the time representing "now" should be selected as 2024 or 2023. Furthermore, perhaps adding a sliding sequence of global averages over a 25-year window would be more convincing. This can also demonstrate the robustness of the research results on a temporal scale.
L183: Given that the author compares the changes in coupling strength here, I believe that a spatial distribution map of the changes in coupling strength should be added to the figure, even though it has already appeared in Figure 2.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have thoroughly addressed my concerns.

(Remarks on code availability)
README code was added, as I suggested in the first round of reviews.

Reviewer #3

(Remarks to the Author)
Thanks for the authors to address my comments. It looks much better than the previous version.There are two further ones:
1)I appreciate the author's description of the GEV for calculting the 2-year and 200 year return periods. As another reviewer pointed out, it may be inappropriate to label the events with a 2-year recurrence interval as "moderate" extreme events. But from my perspective, I believe that a 200 year recurrence interval event as an "extremely" strong high-temperature event is more inappropriate because it rarely occurs. Is it meaningful to study it?
2)Further， it seems that there may be too many Figures for Nature Communications (usually 4-5 charts), and the “Introduction” part of the MS is a bit too long.

(Remarks on code availability)

Response to Reviewer Comments

We thank all reviewers for their encouraging and constructive comments. We addressed all these comments. Please find below a summary of the main changes as well as detailed one-by-one responses to all points raised by the reviewers.

Major changes:

- Most importantly, we added a comprehensive uncertainty analysis. Note that we had assessed model uncertainties by considering the three available SMILES to the extent currently possible (see the discussion below on available SMILES). Our new analysis addresses sampling uncertainty and internal variability, i.e., the statistical robustness of our results. Please note that the 50-member ensemble provides sufficiently many data points to robustly estimate high return levels (as was already assessed for selected locations in Figure 7), changes and correlations. In fact, in some cases, significance tests would not have been very meaningful as even small changes were highly significant. In these cases we used an assessment of changes relative to interannual variability. In other cases we conducted significance tests or calculated confidence intervals. To explain these assessments, we added a long subsection to the Methods section and brief explanations in the relevant figure captions.

- We expanded the analyses to more return periods between 2 and 200 years. For selected locations, we presented such an analysis already in Figure 7, but we now added maps of return level change differences (compared to 2-year return level changes) for 5, 10, 20, 50 and 100 years. These 15 panels have been added to the supplementary information, new Figure S3.

- We found a bug in the calculation of summer mean coupling strength. As a result we needed to correct panels (a) and (b) in Figure 3 as well as Figure S5 and the corresponding discussion in the text. The patterns of significant summer mean coupling shifted a bit polewards, which is even more consistent with results from other studies. These changes did not affect our main results as these are related to changes in very extreme heat and event coupling.

- We reorganised, rephrased and expanded the discussion section to accommodate for the reviewer’s concerns. Additionally, we sharpened the text in some cases by replacing words or rephrasing individual statements. These changes are all highlighted in the document with track changes.

In the following, all line numbers refer to the document with track changes.

Reviewer #1

> *The paper would benefit from a more detailed explanation of the methodology, particularly the calculation of soil moisture-temperature coupling and return periods, to enhance clarity and reproducibility.*

We addressed these points now in more depth in the Methods section.

> *1. My major concern is about the method to calculate return period.*
> *The authors combine all the ensemble realizations (30) and have one return period (30x25=750). I think each simulation should be treated independently and separately when calculating the return period. Each simulation is independent and not a continuous 750-year simulations. I suggest calculating the return periods for each simulation and analyze the mean and variance of those return periods across all simulations. For example, the variances could be used for statical significance of Figure 2. As shown in Figure7, the return value differences can be very different across ensemble members.*
> *Please provide some literatures if the method is commonly used.*

Note that a return level is a property of the climate system, not a property of a single realisation. Return value estimates from individual realisations might be rather different, depending on the internal variability, but this is an expression of sampling uncertainty and not indicating that the underlying population (or true) return levels would differ from sample to sample. They depend only on the underlying state-of-the climate system, i.e., the forcing.
In practice, one only has individual realisations to estimate return levels, therefore our approach might seem somewhat unintuitive. But in fact it is fully rooted in extreme value theory and an elegant and natural way to estimate return levels (note that there is not much literature yet on using SMILES for such applications; Suarez-Gutierrez et al. (Clim. Dynam. 2020) do not precisely explain how their estimation is implemented, and Bevacqua et al. (Nat. Commun. 2023) follow the approach suggested by the reviewer).
Extreme value theory deals with limiting distributions for infinitely long block lengths (Coles, 2001). Using, e.g., annual blocks is a mere practical approximation because of the limited length of available time series and the need to sample many maxima. The assumption underlying the estimation from finite blocks is that the length is sufficiently high and that the block maxima can be assumed to be sampled from the same distribution. This is the case for blocks from different initial condition simulations with identical forcings. There is no theoretical argument against this assumption.
Our approach also allows for the estimation of uncertainty estimates in a natural way following standard statistical theory. Assuming normally distributed uncertainties, one can, by Gaussian error propagation of the standard errors of the GEV parameters, derive approximate 95% confidence intervals (Coles, 2001). An alternative, more sophisticated approach, would be based on the profile likelihood (Coles, 2001).
To demonstrate the validity of our approach, we compare it to the procedure suggested by the reviewer (see Figure below). The results are virtually indistinguishable (Only in locations where the GEV does not properly describe the data, likely because of implausible numerical values in single ensemble members, the method suggested by the reviewer smoothes out the misfit for a single member by averaging over all members).
We prefer using our approach as it estimates the GEV parameters in one go from the full ensemble and is the natural choice when considering extreme value theory.

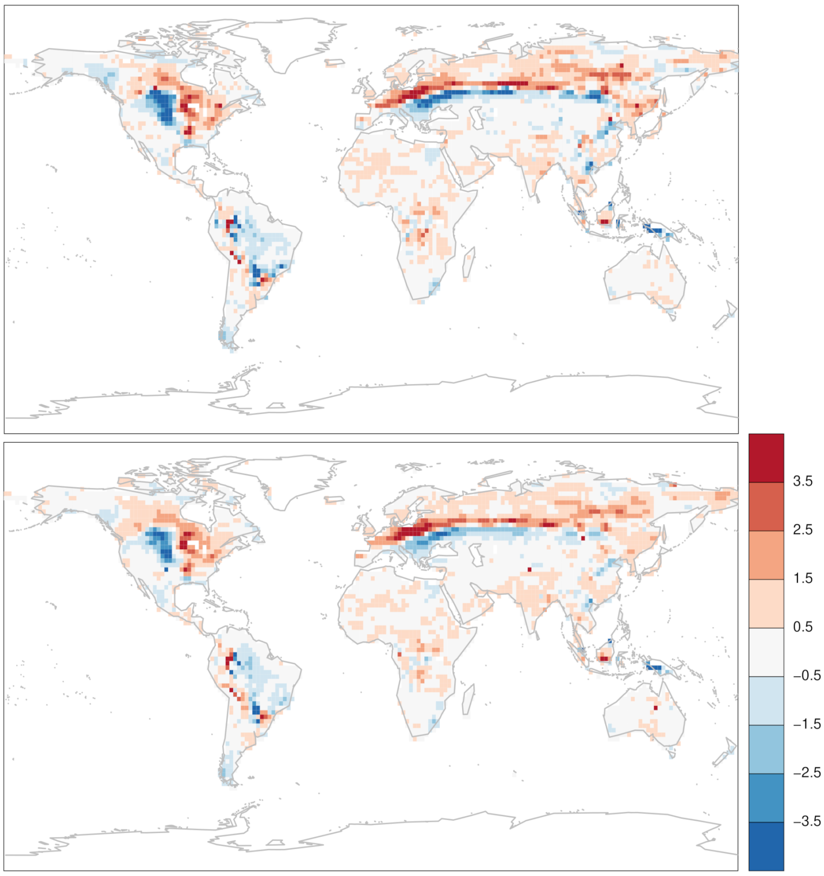

Figure: Comparison of different approaches to calculate differences in 200- and 2-year return level changes simulated by the MPI SMILE. Top: return levels estimated as means over all ensemble members (the reviewer's approach). Bottom: return levels estimated from the full ensemble (our approach). The bottom panel corresponds to Figure 2 (a) in the paper.

> *2. The method of calculating coupling strength is unclear. Specifically (1) Is it correlation between Tx and latent heat or soil moisture? Is the seasonal cycle removed? Please provide equation and examples for positive/negative coupling. How Tx-ET strength relate to soil moisture-temperature coupling for general audiences of Nature Communications (i.e., why not directly use soil moisture-tempearature).*

We are sorry for the confusion about the method of calculating coupling strength as the description in the main text is not consistent with the actual version as explained in the Methods section. It is the correlation between latent heat flux and daily maximum temperature. We have corrected this in the main text.
We compared calculations based on this definition with soil moisture based calculations. The results are qualitatively very similar. As in the literature, the coupling strength is usually derived based on latent-heat flux and temperature, we decided to follow this definition.
For seasonal mean coupling, we consider averages across the summer season in the Extra-Tropics and the whole year in the Tropics (within 20° North/South). Note that we have replaced "warm season" by "summer". For the event coupling we consider values on the hottest day of the year.

> *(2) For correlation across maximum temperature in each year, I think the naming "event" coupling is misleading. It seems more like cross-events relationships.*

We have actually been thinking a lot about an appropriate terminology and agree that "event coupling" is not perfect. But we are afraid that "cross-event" may be misunderstood as well (e.g. as temperature vs. latent heat flux). Our rationale for choosing the term was to distinguish it from seasonal means and to highlight that it considers specific events, namely the hottest day of a year. Another term we had in mind was "episodic", but we were not happy either. We therefore finally decided to keep the term "event coupling".

3.What’s the definition of “total soil moisture content” (Figure 5 b,c)? Is it root-zone soil moisture or total soil moisture in all soil layers? Does deep soil moisture reflect to the occurrence atmospheric surface conditions, or does it reflect to the previous day-to-season weather or hydrological changes (ice, groundwater).

Total soil moisture content is defined as the sum across all soil layers in the land surface model of the corresponding Earth System Model. We did not assess how fast soil moisture responds to atmospheric drivers at different depths, but the fact that seasonal mean precipitation is a driver of coupling changes suggests that the total event coupling depends on a long integration of atmospheric variability, i.e., is affected by soil moisture in considerable depth. Note that standard climate models do not represent groundwater.

4.Additionally, addressing the regional uncertainties and providing more context on the implications for climate risk assessments and adaptation strategies would strengthen the overall impact and relevance of the findings.

Thank you for this suggestion. We have now added several discussion points related to climate risk in the 3rd and 2nd last paragraphs of the discussion. These include a likely lack of preparedness (in the paragraph “But the concerning finding…”), suggestions for better sampling or reducing uncertainties and a discussion of how regional statements following the precautionary principle can still be made based on the physical drivers we identified (in the end of the paragraph “Unfortunately, our assessment…”).

1.Please ensure consistent use of return time(s) or period(s), also return value(s) or level(s).

We have changed to a consistent use of return period and return level.

2.Please specify the raw spatial resolution of the models and the regridded resolution used in this study.

We have added this information in the Methods section. Note that all results are presented on the native resolution of each model (because we did not combine information across models, but considered each model as a self consistent storyline).

3.Please make sure refer to figures, and tables, including supplementary information explicitly in the context.

We checked that we refer now to each figure and table in the main text or relevant figure caption in the main manuscript. A reference to figure S4 (now S5) was missing and has been added.

4.Line 87 Project changes “of”-> in

We have corrected this.

5.For figure 3, please add labels for x- y- axes to the plots for clarity.

We added labels.

6.Line 270-272 realisations-> realizations?

We decided to use British English. This may be adjusted by the copy editors.

7.Line 278: For the warm-season means, how about area in non-tropics?

We made this point clearer and changed the notation. We replaced “warm season” by “summer”, added definitions (JJA in the NH, DJF in the SH). We also added that we consider the whole year in the Tropics (20° North to 20° South).

8.Check if Txx is a typo of Tx.

Thank you for spotting this. Txx is the variable name for the annual maximum of Tx. We actually avoided using Txx in the text (and rather referred to annual maxima of Tx). We now explicitly introduce this notation in the methods section and the main text.

Reviewer #2

> *Overall, I think this is a very interesting study, and the methods are straightforward and justifiable. I have some suggestions that I recommend the authors address prior to publication, as outlined below:*

We thank the reviewer for their positive assessment. We are confident that we have addressed all the reviewer's concerns.

> *[1] I think it would be nice to more explicitly describe what constitutes an extreme event using return periods. I question whether a 2-year event is considered "extreme", as it seems more like normal, frequent event. Is an event with a 2-year interval really a "moderate" extreme event? Furthermore, why do the authors choose to compare a 2-year event to 200-year events? Why not along a gradient of return periods? Analyzing just the end points seems like a missed opportunity.*

We agree with the reviewer that a 2-year event is not really an extreme event – this is a key point of our paper. Early in the introduction we discuss the mismatch between impactful events, which often have return periods of hundreds of years, and the typically considered extreme event indicators which often refer to events happening even several times per year (e.g. the 95th percentile of daily values is exceeded every 20 days on average, even the 99th percentile is exceeded some 3-4 times per year). We chose the 2-year return period to address such events. When introducing 2-year events as moderate extremes, we refer to the preceding discussion on standard extreme event indicators (page 2, line 74).
We agree with the reviewer that considering a continuum of return periods would add valuable information. In fact, figure 7 shows such a continuum for two selected locations. Additionally, we now added additional figures for 5, 10, 20, 50 and 100-year events. Given the huge number of panels, we had to move these plots to the supplementary information, Figure S3.

> *[2] L135-145, I am a bit confused how moving from a transitional to dry regime would dampen a heatwave. Even though the coupling is strongest in transitional regimes, this does not imply that heatwaves are amplified in transitional regimes, right? If a transitional regime, for example, became water-limited, all the available energy would go to sensible heat, heating the surface and increasing near-surface air temperature more than if there was any soil moisture available.*

We thank the reviewer for spotting this issue. Of course, drying out reduces latent heat fluxes, increases sensible heat fluxes, and thereby increases the maximum temperature during a heatwave, as during the transition from a transitional to a dry (water-limited) regime. The point we actually wanted to make was rather about the amount of amplification during the heatwave, not absolute temperatures. We deleted the corresponding statements from the manuscript.

> *It might benefit the paper to think about the distinction between transitional regimes in space vs. time.*

We are not quite sure we correctly understand the reviewer's comment.
In the classical literature (e.g. Seneviratne et al., 2010), soil moisture temperature coupling was implicitly assumed to be constant in time. In this context, the term "transitional" refers to a state with intermediate soil moisture values (low enough not to be energy limited, high enough to allow for coupling). The temporal aspect of coupling strength (and therefore possible regime changes in time) as discussed recently by Hsu & Dirmeyer (2023) is a key starting point for our paper as

discussed above. Our main aim is to assess long term climatic changes in the coupling strength on the hottest days of a year.
Spatial aspects could also be discussed from two perspectives: we present the spatial patterns of coupling regimes and how they change in a warming climate. Another perspective could be to discuss “spatial transitions” as e.g. the spatial propagation of soil moisture drought during individual events. This aspect, however, is beyond the scope of our paper and merits separate studies. Given the limited space, we therefore decided not to mention this point.
Related to this discussion, note that the term “transition” here refers to two issues: (1) in the traditional concept, the “intermediate” regime where coupling is possible; and (2) to the transition between regimes. Given this new notion of regime transitions, the former, older, term “transitional” appears to be somewhat misleading. We decided to still use it, as it is common in the literature, and instead replaced the verb “transition” (as from wet to transitional) by “change” in the text.

> *Also, it could be interesting to include a conceptual figure using example data showing how event coupling is calculated in the SI.*

We now better explain the calculation of event coupling in the methods section. As the actual calculation as the correlation of Txx and LH (at the occurrence day of Txx) is rather simple, we believe it is not necessary to include an example calculation or an additional conceptual figure.

> *[3] It is unclear what is considered the “warm season”, and if seasonality during the warm season affects the coupling strength, i.e. is the seasonality driven by seasonality or daily anomalies?*

We replaced “warm season” with “summer” with the additional explanation that we use annual values (means, maxima) in the Tropics. It is likely that seasonality during the summer influences the coupling strength, as the warmest values would cluster in a certain time of the year and the typical precipitation during that day would again depend on the time of the year. This seasonal dependence, however, does not have any implications for our methodology as the choice of the maximum values within a year intrinsically accounts for this seasonality. We agree, however, that understanding how much day-to-day variability and seasonality affect the timing of the hottest day in a year for a given region would be an interesting study in itself.

> *[4] Figure 4, The arrows below the conceptual figure are confusing. Do the arrows correspond with the conceptual figure above? More importantly, though, I am not sure this figure adds much to the paper, as the information seems to be easily described in text.*

We actually believe the visualisation strength of the figure lies in the combination of the three regimes with the arrows. The arrows are consistent with the conceptual figure and describe changes in coupling regimes from present to future climate. We clarified this in the figure itself and added to the figure caption that these 4 examples are discussed in the main text (in the discussion of Figure 5). The Eastern US were not mentioned in the text, so we added them. Vice versa, we added a reference to Figure 4 (the conceptual figure) when discussing the examples in the text. We thought about adding circles or crosses at these locations in the maps, but decided against this as such icons would either be hardly visible or hide too much information in the figures.

> *[1] Introduction: I think the readability would improve if the one sentence paragraphs in the introduction were incorporated into larger paragraph.*

We combined paragraphs accordingly.

[2] has to have

We changed this.

[3] L61, These seem more like the key results, not your hypothesis.

We agree that these points emerge as our key results based on the analyses presented in the study. However, in the introduction we present these as hypotheses to motivate the study.

[4] L84, "of our results" should be more specific

We specified this in the text (and added also that storylines are physically self-consistent).

[5] L85, check grammar

We added the missing "we".

[6] The language in the last four paragraphs feels too casual and vague, "good news", "bad news", "huge model uncertainties", "But still".

We adjusted the language.

It appears that there is code to create each figure. There is no README file describing each file, though.

We replaced the old code with the new one generating the uncertainty analysis and also added a README file describing the indicidual scripts. Note that there is one file that executes all files and produces all figures.

Reviewer #3

L73: Warm season is not clearly defined in the text, which specific months is it?

We replaced “warm season” by “summer”, i.e., JJA in the NH and DJF in the SH, with a remark that we considered the whole year in the Tropics.

L80: Why use these three large sample models? Have you considered other large sample models? The physical process simulations between different models are inconsistent and may result in inconsistent results.

We wanted to analyse as many SMILES as possible. For our analyses we chose all SMILES that provided the required variables at a daily resolution for a common (medium or high-emission) scenario. For comparison, Olonscheck et al. (2023) provide an overview of available CMIP6-based SMILES: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023MS003790. Most of these seven ensembles did not provide the required data. We specified this more explicitly in the methods section.

L88: The article does not explain why SSP585 is used to continue the results of the model after 2014. Will the response of extreme heat events to soil moisture temperature coupling be inconsistent under lower emission scenarios?

As indicated above, the choice was more or less enforced by the available ensembles. An alternative would have been the SSP1-2.6, but we considered this being too optimistic given our focus on climate risks (and thus the need to explore at least medium-high emission scenarios). We are of course aware of the fact that SSP5-85 is a very pessimistic worst-case scenario. Therefore we also considered mid-century projections (Figure 6), where SSP5-85 is still close to other scenarios. The fact that the same pattern as shown for the end of the century emerges, demonstrates that our results are also relevant for both lower emission scenarios and mid-term projections. This has been explicitly stated already in the initial submission (page 6, line 232 – we slightly rephrased the statement).

Furthermore, I believe that the period representing the 'present' should be chosen to be 2024 or 2023, rather than 2014 when the Historical scenario ends. The SSP scenario can be used to extend the subsequent years.

Our intention was not to provide an assessment relative to exactly present climate, but to demonstrate the physical effect and its relevance. For this purpose, the choice of reference (as well as future) periods is somewhat irrelevant as long as it considers (a) a range with a strong signal representing high-emissions and/or long-term changes and (b) a range with a moderate signal representing medium emissions and/or mid-term term changes. We therefore decided to keep the references period. We considered renaming this period “historical” or “recent”, but found both terms potentially misleading (historical implying a distant past, recent implying the last couple of years).

And, there is no uncertainty-related information in the graph, do all grid results have a high level of confidence? I hope the author takes into account the issue of uncertainty when analyzing.

In our context, the most relevant uncertainties are climate response uncertainty and forcing uncertainty (see IPCC AR6 WG1 Chapter 1, Chen et al., 2021 for definitions). Both have been covered by (a) comparing different climate models and (b) by comparing two projection horizons (mid/end-of century). Note that, for climate response uncertainty, we deliberately decided NOT to present multi-model ensemble means as the identified spatial patterns were qualitatively so different that averaging would not have made any sense. Instead, as discussed in the introduction, we considered each model independently as a plausible physically self-consistent storyline (Shepherd et al., 2018, and IPCC AR6 WG1, Chapter 10, Doblas-Reyes et al.).
The third source of uncertainty, internal variability and thus sampling, is of much lower relevance as we use SMILES with many realisations. However, we agree with the reviewer that it is not a priori clear how robust our estimates of high return levels actually are, even when based on SMILES. We thus assessed the sampling uncertainties and role of internal variability for our results in the revision.
Here we would like to highlight an important point. Given the high number of ensemble members, there is a danger that even tiny changes or differences are significant, highlighting that significant results are not necessarily relevant when the sample size is large enough (The famous statistician Fisher summarised this dilemma as “All Null hypotheses are wrong”). In fact we found exactly this result for most of the aspects considered. Therefore, we decided to deliberately avoid significance tests where possible and rather asked the question “Are the detected differences/changes relevant?”. We considered “relevant” as being large compared to a chosen fraction of interannual variability. For some aspects, however, we did consider significance tests as interannual variability could not easily be defined. These are (1) the coupling strength in present climate; (2) the changes in coupling strength; and (3) differences in changes of return values. We added a new sub-section to the methods section to discuss and explain these issues in depth, brief explanations in the discussions of the results, and short statements in the corresponding figure captions.

> *L94: To characterize the correlation between two variables, confidence analysis should be added instead of just presenting correlation data. In addition, almost all the data in the article does not include uncertainty ranges, which has caused me great concern.*

As discussed above, we added significance tests. Our key results are all highly significant.

> *L134: What conclusions did the seasonal average coupling strength of these models reproduce from previous studies? There is no necessary explanation for Figure S3, making it difficult for readers to understand its meaning.*

We added brief explanations to the text and the caption of Figure S3 (now S4, we also refer to S5 now) and refer to two relevant papers. We also make it clear now that S5 refers to the amplification of changes in moderate extremes, and state that the actual spatial patterns are displaced across models (something which is typically hidden in the multi model mean plots of other studies).

> *L140: This is very counterintuitive. Humidity should always have a restraining effect on temperature drop, as it reduces the evaporative cooling effect on the surface, leading to an increase in temperature. Although the strength of land-atmosphere coupling decreases, the temperature should be increasing. I hope the author can provide clearer explanations.*

We thank the reviewer for spotting this issue. Of course, drying out reduces latent heat fluxes, increases sensible heat fluxes, and thereby increases the maximum temperature during a heatwave, as during the transition from a transitional to a dry (water-limited) regime. The point we actually wanted to make was rather about the amount of amplification during the heatwave, not absolute temperatures. We deleted the corresponding statements from the manuscript.

L146: As above, the uncertainty range should be added to the graph, and the time representing "now" should be selected as 2024 or 2023. Furthermore, perhaps adding a sliding sequence of global averages over a 25-year window would be more convincing. This can also demonstrate the robustness of the research results on a temporal scale.

We added confidence intervals to the figure. Note that these confidence intervals are deliberately chosen as conservative upper bounds: The estimation of the degrees of freedom of zonal averages over land areas is not trivial, so we assumed that for each latitude we had only one data point. The true confidence intervals are thus much narrower, depending on the landmass and spatial dependencies at any given latitudinal band.
As discussed above, we decided not to change the reference period. The spatial patterns are highly regional, such that a global average sliding sequence would average out all interesting features.

L183: Given that the author compares the changes in coupling strength here, I believe that a spatial distribution map of the changes in coupling strength should be added to the figure, even though it has already appeared in Figure 2

We are afraid that repeating this information would be difficult. Adding contour lines to panel (c) would make this map overloaded with information. We additionally double checked with the editor that Nature Communications would not allow repeating figure panels.

Response to remaining Reviewer Comments

Reviewer #3

> *1)I appreciate the author's description of the GEV for calculting the 2-year and 200 year return periods. As another reviewer pointed out, it may be inappropriate to label the events with a 2-year recurrence interval as "moderate" extreme events. But from my perspective, I believe that a 200 year recurrence interval event as an "extremely" strong high-temperature event is more inappropriate because it rarely occurs. Is it meaningful to study it?*

As discussed in the introduction of our manuscript, the highest impacts typically arise from very rare exteme events such as in the European heatwave from 2003, the Russian heatwave from 2010 or the Pacific North West heatwave from 2021. For climate risk assessments, these events are therefore highly relevant. In fact, it has been prominently argued that climate risk assessments should also address the highest risks (e.g. Weaver et al., Env. Res. Lett., 2017; Sutton, Bull. Amer. Meteorol. Soc., 2019). Studying how these events change in a warming climate is thus important. Currently, research on such events is limited, and assessments are typically built on the assumption that very extreme events change similar to moderate extremes. Here, we demonstrate that this assumption is wrong for large regions of the globe.

> *2)Further, it seems that there may be too many Figures for Nature Communications (usually 4-5 charts), and the "Introduction" part of the MS is a bit too long.*

We double checked that the number of display items and the length of the introduction are in line with Nature Communications formatting requirements. Nevertheless, we slightly shortened the introduction.