Peer Review File

Rarest rainfall events will see the greatest relative increase in magnitude under future climate change

Web links to the author's journal account have been redacted from the decision letters as indicated to maintain confidentiality.

28th Mar 22

Dear Dr van der Ent,

First of all, please allow us to apologise for the delay in sending a decision on your manuscript titled "Rarest rainfall events have greatest relative increase under climate change". It has now been seen by 2 reviewers, and I include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter) and the completed checklist:
[link redacted]
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to receive your revised paper within six weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

We understand that due to the current global situation, the time required for revision may be longer than usual. We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning. Of course, if you are unable to estimate, we are happy to accommodate necessary extensions nevertheless.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Clara Orbe, PhD
Editorial Board Member
Communications Earth & Environment
orcid.org/0000-0002-5344-3599

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded as a Related Manuscript file type with the revised article.

Editorial Policy: Policy requirements

Furthermore, please align your manuscript with our format requirements, which are summarized on the following checklist:
Communications Earth & Environment formatting checklist

and also in our style and formatting guide Communications Earth & Environment formatting guide .

*** DATA: Communications Earth & Environment endorses the principles of the Enabling FAIR data project (http://www.copdess.org/enabling-fair-data-project/). We ask authors to make the data that support their conclusions available in permanent, publically accessible data repositories. (Please contact the editor if you are unable to make your data available).

All Communications Earth & Environment manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

In particular, the Data availability statement should include:
- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: All new data associated with the paper should be placed in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at http://www.nature.com/sdata/policies/repositories.

If a community resource is unavailable, data can be submitted to generalist repositories such as figshare or Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent

URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at http://www.nature.com/authors/policies/availability.html.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The paper looks at changes in daily rainfall extremes in 25 recent CMIP6 model simulations and compares changes to the yearly extremes to those occurring once every (10) 100 years. They find that changes to the latter "rare" extremes are larger (when expressed in relative terms). This finding is interesting and very relevant, and is a timely contribution to the field. The finding that the most extreme events increase the most is not entirely new, but it is a very important message and, as far as I know, the first time this question has been studied in such a great depth. Therefore, I recommend publication. But I also have some important point to consider as detailed below.

Major:

*
the number provided are not the same between discussion and abstract. Figure 3 appears to support the numbers provided in the discussion. Would be good to check all numbers in the document.

*
weighting of model results is clearly a valid approach. But looking at the Table 2, it looks like there is quite a wide range in weights and I am not sure this can really be defended based on the criteria. Therefore, it would be good to at least also provide some "unweighted" estimates, or investigate whether there is a correlation between the weights and the response, or provide some information how weighting affected the results.

*
Although the paper focuses on the statistics, it would still be nice if the authors could speculate about the cause of this behaviour. Could it be related to dynamical feedback mechanisms due to latent heating as discussed e.g. in Nie et al. Or perhaps changes to the frequency of rain play a dominant role. In areas where the frequency of rain events decreases it is usually observed that the most extreme events change the most. (this is already obtained in a simple conceptual model with a exponential tail of the rainfall distribution, combined with constant fractional change of the wet conditional distribution and a decrease in rain frequency without the need for dynamical amplification). It looks like that the dryer areas in Table 2 (or areas that get dryer in the future) are also the large amplification areas.

*
I could not find information on how the averages are computed, and how the weights are actually used. There are is quite some averaging involved here: averages across regions and models, and would be good to spell out how this is done. Doing regional averages it could be advantageous to do a geometric average since in a noiseydata sets the normal mean could introduce a positive bias in

the response; see Appendix of Aalbers et al. 2018, it can be shown that if the error is multiplicative, so P*(1+ eps), a bias in the response is introduced equal to eps**2 using the arithmetic average (but none in a geometric mean).
*
The introduction could be improve by better embedding in existing literature. Some papers discuss the larger changes to the more rare extremes (like Aalbers et al, Hodnebrog et al,) and I guess there are more. Also Li et al already performed an elaborate analysis of changes in precipitation extremes in CMIP6.

minor:
*
L 23-33. Perhaps would be good to mention work on CMIP6 and precipitation extremes, such as in Li et al (2021). This could be at the expense of the discussion on small-scale/short duration extremes. I am not sure whether you can really say that daily are good enough represented in the CMIP models, considering the that in the tropic these are of convective origin. I'd rather say something like that despite these limitations, GCMs are still the only available tools to look at (rare) extremes at a global scale.
*
L31 it's rather large-scale dynamical or synoptic scale dynamical. Also for small-scale events changes to the dynamics are very important, perhaps even more than for the large-scale events.
*
L42. Extreme value theory are a tool to study these rare extremes, although in principle that can also be derived from single model initial condition large ensembles (SMILEs) or spatial pooling of results. Perhaps formulate this more generally.
*
L97. Would be valid information to also have the change in the 1-year return level in this table

Nie J, Sobel AH, Shaevitz DA, Wang S (2018) Dynamic amplification of extreme precipitation sensitivity. Proc Natl Acad Sci USA 115:9467–9472. https://doi.org/10.1073/pnas.1800357115
Hodnebrog Ø, Marelle L, Alterskjær K, et al (2019) Intensification of summer precipitation with shorter time-scales in Europe. Environmental Research Letters 14:124050. https://doi.org/10.1088/1748-9326/ab549c
Li C, Zwiers F, Zhang X, et al (2021) Changes in Annual Extremes of Daily Temperature and Precipitation in CMIP6 Models. Journal of Climate 34:3441–3460. https://doi.org/10.1175/JCLI-D-19-1013.1
Aalbers EE, Lenderink G, van Meijgaard E, van den Hurk BJJM (2018) Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability? Climate Dynamics 50:4745–4766. https://doi.org/10.1007/s00382-017-3901-9

Reviewer #2 (Remarks to the Author):

In this paper the authors aim to assess if rare extremes (return period 100 years) will relatively more than common extremes corresponding in return periods of 1 year. The question is relevant, but from a statistical point of view, given the assumption that warming will cause changes in precipitation regime, the answer is yes. Assuming a nonstationary distribution expressing this change then the relative change, in general, cannot be the same at T =1 year and T = 100 years, or for different T's in general (I provide details later on). Yet whether it's going to be larger or smaller change in one of the other case depends on the type of nonstationarity. Overall, the paper is well-written and well organized and deserves communication. Next I provide some specific comments that the authors might want to consider.

42 annual maxima are still based on extreme value theory

139 just to be 100% sure by the historical late 20th century period 1971-2000 the authors refer to the simulated historical as mentioned in 52. I would add it again to avoid confusion.

140-141 A general point is why include only the last 30 years of the 21th century. A policy/decision maker might be also interested in the next 30 years maybe even more than 70 years ahead.

142-143 what is the rational of selecting the first of the available simulation of each model? Why not select the best simulation in terms of observations? I mean compare simulation with observed data and select the best performing.

146 some models have quite coarse resolution; I believe remapping at 0.25 x 0.25 could affects the results. At least the authors should discuss these points and mentioned potential issues that might emerge.

Model Weighting. My understanding is that this is a popular weighting scheme that have been adopted and is based on the 5 points mentioned. How these points are really important for precipitation extremes? Wouldn't be more important to test if the model reproduces the actual observed distribution of extremes? Variability is important in precipitation but you can have same variability and different tails for example.

183-185 well this could be the case only if the tail used in MEV matches reality. The Weibull tail shows in general good results, but other studies show also good results for power-type tails. So maybe add this assumption after the "is therefore able to estimate return periods higher than the period of record with reduced uncertainty" if the tail matches. The fact that you might get smoother patterns does not mean that you don't get consistent bias which you'd get as I mentioned if there's a mismatch in the asymptotic behavior.

Statistical Analysis. One could approach this problem a bit differently or use another method to verify the results. The question of whether the rarest extremes will increase more than the common ones can be also replied if one investigates the nonstationary GEV or any other nonstationary distribution. For example if you use the historical simulation + the future simulation and fit a nonstationary gev then you get the answer to your question easily since you can estimate the relative change as 100(Q_GEV(a(t2),b(t2),c(t2),1-1/T)/ Q_GEV(a(t1),b(t1),c(t1),1-1/T)-1) for T =1 year and then for T =100 years and for any times t2 (e.g, 2100) and t1 (e.g., 2000). This will give you the answer. Of course different types of nonstationarities can be assumed here (e.g., changes in location parameter, or scales, or even shape (not advised) or combinations).

A final point of discussion which I might have missed is if there is any validation of the simulations with actual observations. I mean CMIP6 or climate models in general sometimes predict precipitation from another planet, what is the point of the analysis if the models don't predict well extreme precipitation? To clarify, I understand that you wish to investigate relative changes, yet since the projections have not been bias corrected then results shown might be misleading. For example results shown in Fig 2 and 3 or mentioned in the abstract and discussion can give a wrong impression. My understanding is that you have not bias corrected the simulations and have not compared with historical observations. Thus the analysis shows the relative change in the simulation of a model; if the model is bad then such numbers can be misleading. The D measure seems better in this sense.

Overall, I believe this is a decent paper and deserves communication.

Reviewer #1 (Remarks to the Author):

The paper looks at changes in daily rainfall extremes in 25 recent CMIP6 model simulations and compares changes to the yearly extremes to those occurring once every (10) 100 years. They find that changes to the latter "rare" extremes are larger (when expressed in relative terms). This finding is interesting and very relevant, and is a timely contribution to the field. The finding that the most extreme events increase the most is not entirely new, but it is a very important message and, as far as I know, the first time this question has been studied in such a great depth. Therefore, I recommend publication. But I also have some important point to consider as detailed below.

Many thanks to the reviewer for their time, insightful comments, and finding our study interesting, very relevant, and a timely contribution to the field. We have looked at the points that are raised, and respond to each of them below.

Major:
* the number provided are not the same between discussion and abstract. Figure 3 appears to support the numbers provided in the discussion. Would be good to check all numbers in the document.

You are correct to notice the difference in the numbers between the abstract and discussion. In the abstract the numbers represented the change in “daily land rainfall extremes” (L16), whereas in the discussion the numbers reflect the “daily rainfall extremes” (L124), so world-wide instead of only over land. We regret that this was not entirely clear, and have included two tables containing these numbers as the change in T1 as well as the change in T100 estimates as respectively Table S3 and S4. We have also rephrased this in the discussion and added references to these tables L123-126:
“Particularly we found for low emission scenario SSP1-2.6 and high emission scenario SSP5-8.5 global (land and ocean) daily rainfall extremes will increase by 8.6 % and 23.1 %, respectively, for yearly events (Table S3) and by 11.9 % and 32.5 % for for 100-year (Table S4) events by the end of this century.”

* weighting of model results is clearly a valid approach. But looking at the Table 2, it looks like there is quite a wide range in weights and I am not sure this can really be defended based on the criteria. Therefore, it would be good to at least also provide some "unweighted" estimates, or investigate whether there is a correlation between the weights and the response, or provide some information how weighting affected the results.

Thank you for pointing this out. The range of the weights (mostly 0.01 - 0.1; with some exceptions being as low as 0.001) is comparable to similar recent studies, such as in Brunner et al. (2020; their table S2 in the supplement) looking at global temperature changes. In ClimWIP the strength of the weighting is established using a leave-on-out model-as-truth test (also called "perfect model test") as described in detail, for example, in Brunner et al. (2020; see section 2.5 and section S3 in the supplement). Here the target that was optimised was global mean precipitation change between two periods: 1995-2014 to 2080-2099. In essence the calibration of the weights using this method means that the weighted distribution is perfectly reliable by definition in predicting out-of-sample "perfect models". In order to explain this better, we have expanded the Model weighting section L186-200 in the methods with the following:
“For model performance we adapted the metrics used in [35] to the target of global precipitation change. In contrast to other important climate variables, most prominently future warming [35, 55, 56, 57], emergent constraints [58] for global precipitation changes have only recently been suggested [59, 60]. Here, our main aim is to reduce the influence of models which simulate variables considered

important for the representation of precipitation very different from observations rather than applying a constraint that necessarily reduces model spread. Performance weights were, therefore, based on five metrics: 1) the temperature trend, which has been found to be an important constraint for temperature and precipitation changes alike [55, 59], 2) the temperature climatology, 3) the variability of temperature, 4) the precipitation climatology and 5) the variability of precipitation, all in the period 1979-2014. Models which perform badly in one or more of these metrics receive less weight in the calculation of multi-model statistics as we trust their projections of future precipitation change less. The strength of the weighting is established using a leave-one-out model-as-truth test [37, 61, 35] on the target of global mean precipitation change. The resulting weights for each model are included in Table 2 and have a range comparable to recent studies, such as Brunner et al. [35] (their table S2 in the supplement). The weights of the models were used to create the multi-model weighted ensemble means for figures 1, 2, 4, S3, S7-S9, and S23-S25."

We agree that it is also important to demonstrate how the weights influence the results. In order to do this, we have included the unweighted estimates in the supplement (Figures S1-S3). As these unweighted figues show, the weighting has minimal impact on the results. We added the following text to the manuscript L61-64: *"The multi-model ensemble means that are shown in this manuscript are weighted for their independence and performance, but we note that using unweighted estimates does not considerably affect the results and conclusions drawn in this study (see Figs. S1-S3 in the supporting information)."*

* Although the paper focuses on the statistics, it would still be nice if the authors could speculate about the cause of this behaviour. Could it be related to dynamical feedback mechanisms due to latent heating as discussed e.g. in Nie et al. Or perhaps changes to the frequency of rain play a dominant role. In areas where the frequency of rain events decreases it is usually observed that the most extreme events change the most. (this is already obtained in a simple conceptual model with an exponential tail of the rainfall distribution, combined with constant fractional change of the wet conditional distribution and a decrease in rain frequency without the need for dynamical amplification). It looks like that the dryer areas in Table 2 (or areas that get dryer in the future) are also the large amplification areas.

This is indeed an interesting question. Using the Weibull distribution (underlying the MEV-distribution) it can indeed conceptually be shown that the phenomenon of the rarest extremes increasing more can either be caused by a decrease in the number of wet days N (with an increase of the scale parameter C in order to still end up with a net increase) or a decrease in the Weibull-shape parameter W. Without going into disentangling this in detail we have now provided spatial figures in the Supporting Information (Figures S23-S25) showing the changes in all three parameters of the Weibull distribution. From those figures it can be observed that both processes play a role. We have now included the text below in the discussion L138-145:
"Yet, when looking at the changes in the parameters underlying the MEV-Weibull distribution (Equation 4), the statistics itself give some indication about the processes (Figures S23-S25). It should be noted that the behaviour could be caused be either a decrease in the number of wet days N (combined with an increase in the scale parameter C) [48] or a decrease in the shape parameter W, which may, for example, be the result of dynamical feedback processes related to latent heating [49]. It appears that both the rainfall frequency (Figure S23) as well as dynamical feedback processes (Figure S25) play a role. This may serve as a starting point for future research to further disentangle the processes behind this behaviour."

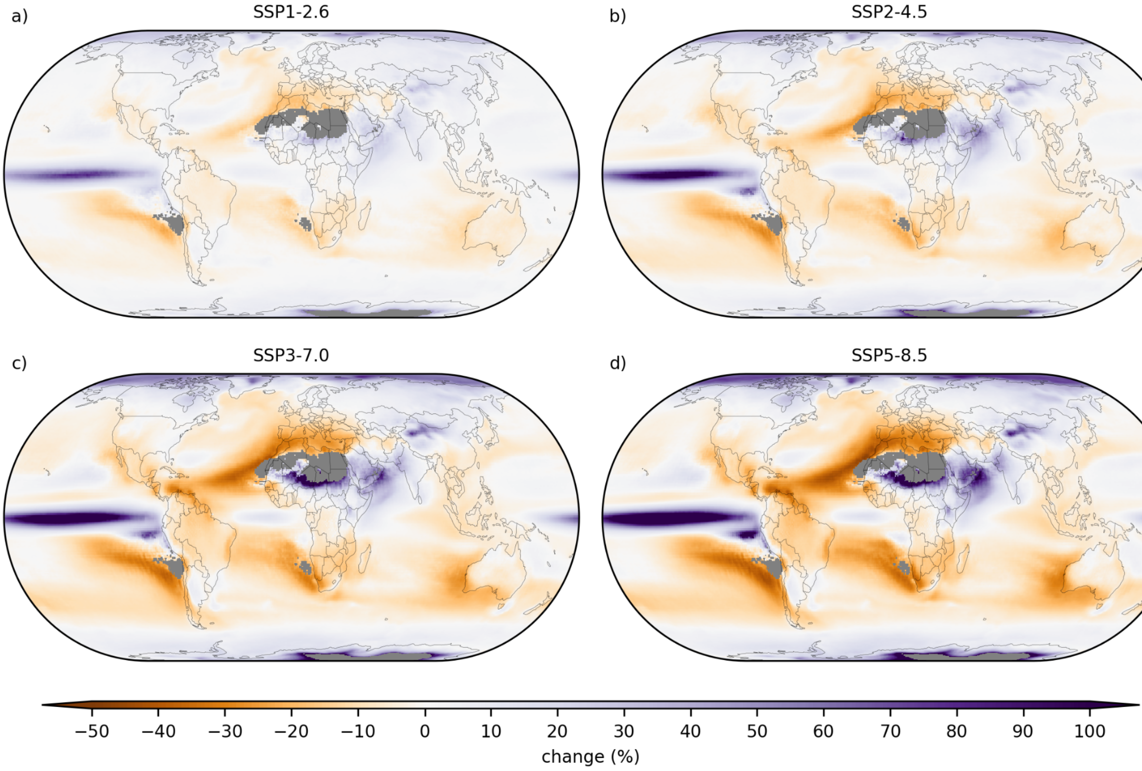

Figure S23: Relative change of the mean number of wet days per year (N parameter for the MEV distribution) expressed as the weighted mean across 25 CMIP6 models for the (a) 1-2.6, (b) 2-4.5, (c) 3-7.0 and (d) 5-8.5 future periods (2071-2100) with respect to the historical period (1971-2000) (eq. 2). Hatching is in the locations where >75 % of the models agree on the sign of the change. Dry areas (weighted mean of less than 3 events per year) are masked in grey.

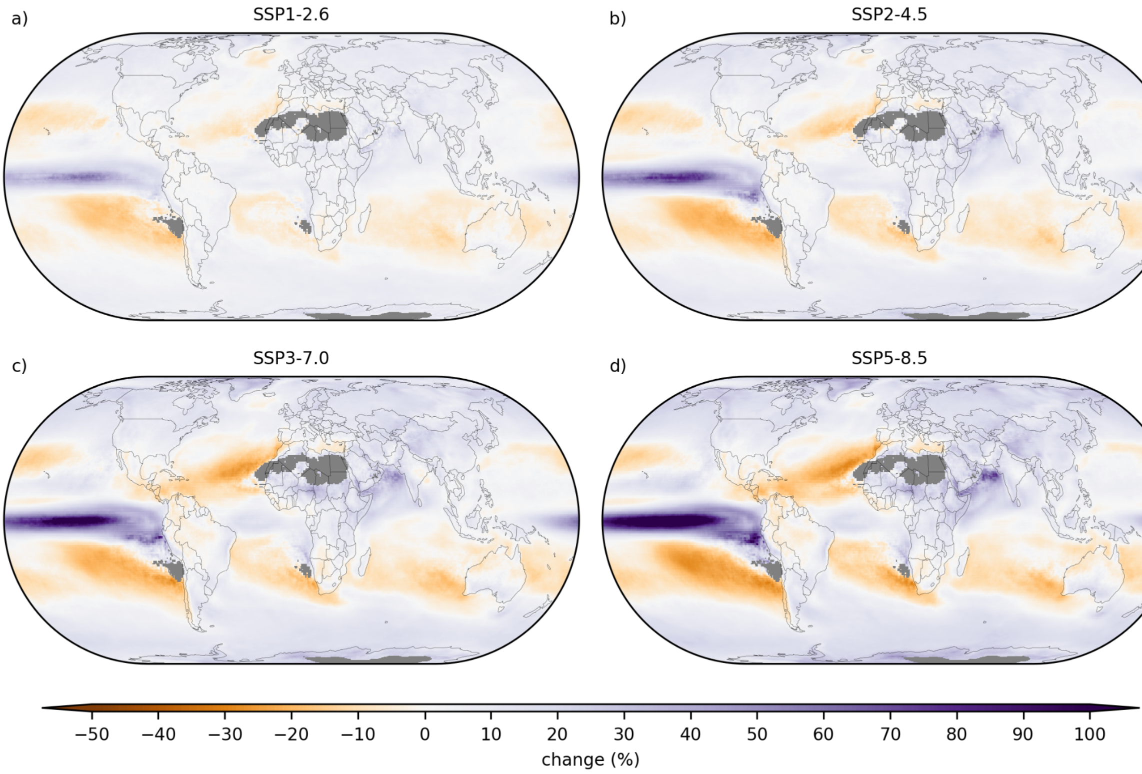

Figure S24: Relative change of the mean Weibull scale parameter C, expressed as the weighted mean across 25 CMIP6 models for the (a) 1-2.6, (b) 2-4.5, (c) 3-7.0 and (d) 5-8.5 future periods (2071-2100) with respect to the historical period (1971-2000) (eq. 2). Hatching is in the locations where >75 % of the models agree on the sign of the change. Dry areas (weighted mean of less than 3 events per year) are masked in grey.

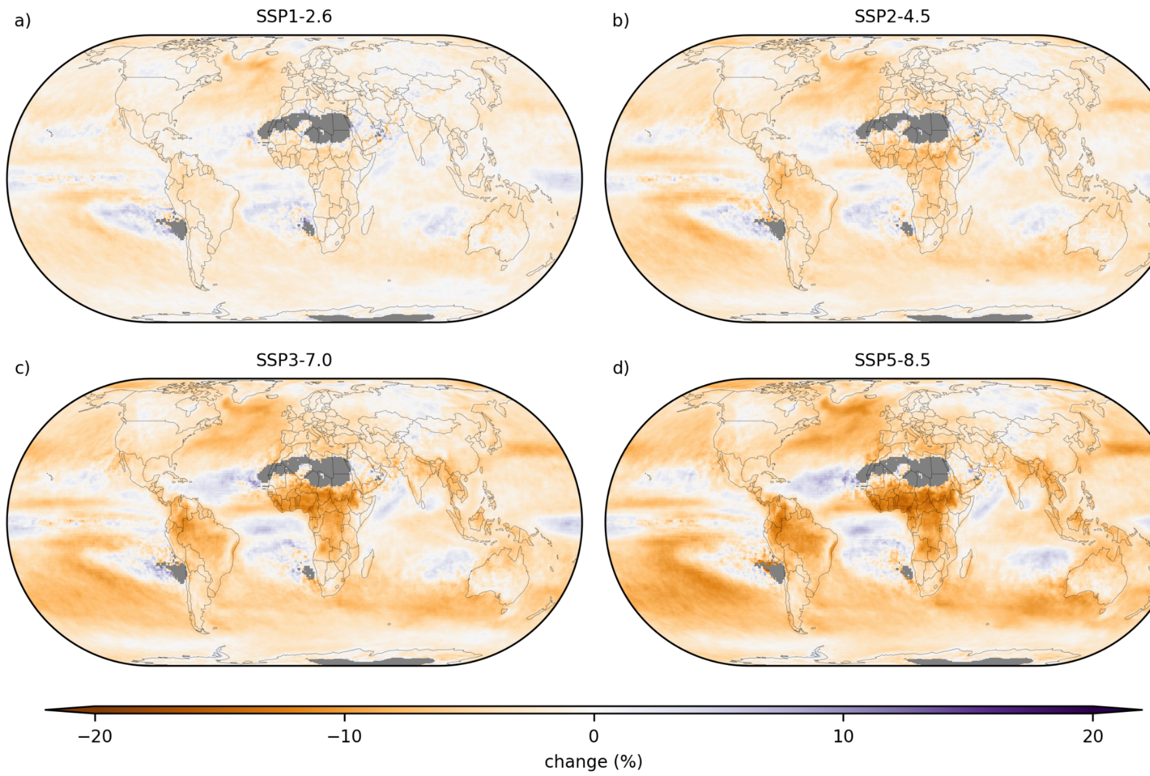

Figure S25: Relative change of the mean Weibull shape parameter *W,* expressed as the weighted mean across 25 CMIP6 models for the (a) 1-2.6, (b) 2-4.5, (c) 3-7.0 and (d) 5-8.5 future periods (2071-2100) with respect to the historical period (1971-2000) (eq. 2). Hatching is in the locations where >75 % of the models agree on the sign of the change. Dry areas (weighted mean of less than 3 events per year) are masked in grey. Lower *W* parameter is indicative of a heavier tail of the distribution.

* I could not find information on how the averages are computed, and how the weights are actually used. There are is quite some averaging involved here: averages across regions and models, and would be good to spell out how this is done. Doing regional averages it could be advantageous to do a geometric average since in a noiseydata sets the normal mean could introduce a positive bias in the response; see Appendix of Aalbers et al. 2018, it can be shown that if the error is multiplicative, so P*(1+ eps), a bias in the response is introduced equal to eps**2 using the arithmetic average (but none in a geometric mean).

We regret to hear that the averaging was not entirely clear. We have to admit that we were not very familiar with the use of the geometric vs. the arithmetic mean. Looking at the description in Aalbers et al. (2018), we acknowledge that a geometric mean may sometimes be beneficial to avoid biases. However it also comes with other issues: technical problems when dealing with negative values and it can be less intuitive to interpret for non-statisticians than the commonly used arithmetic mean. Moreover, in Figure 2 we focus on the median and not the mean. Therefore, we chose to stick with the arithmetic mean and median. Nonetheless, we agree that it is important to describe our averaging procedures in more detail. First of all, we did all analyses in the native resolution of the models, e.g., calculating the return levels, calculating the differences in T100 ($C_{rel,T100,x}$, eq. 2) or T1 estimates ($C_{rel,T1,x}$, eq. 2), and calculating the $D_{100-1,x}$ (eq. 3): the difference in the change in 100-year return level ($C_{rel,T100,x}$, eq. 2) minus the change in the 1-year return level ($C_{rel,T1,x}$, eq. 2). After we performed all analyses in the models' respective native resolutions, we then used the nearest neighbour interpolation method to remap the results to a higher resolution (0.25° × 0.25°) grid. The remapped grid cells are 4 to 63 times smaller than the models' native grid cells, and by using nearest neighbour interpolation no spatial averaging has to occur.

In order to create the multi-model ensemble means, we apply the weights for each individual model to the remapped results to create figures 1, 2, 4, S3, S7-S9, and S23-S25. For regional or global averaging (tables 1, S3 and S4) we also included information on the cell size by giving higher weights to the cells closer to the equator than to the poles.

To clarify this further, we rephrased or added sentences to several sections in the method section. To the "CMIP6 model data" L167-174 we added: *"As there are large differences in the resolution of the different GCMs, the analyses are performed on each model's native grid. The results are then remapped to a 0.25° × 0.25° grid using the nearest neighbour interpolation method for the ensemble means. The results remain mostly unaffected by the remapping, as the 0.25° × 0.25° grid is a higher resolution than the native grid of each of the models. Specifically, the remapped grid cells are 4 to 63 times smaller depending on the model resolution, and as we used nearest neighbour interpolation no spatial averaging is taking place. Moreover, as the main results of this study are relative values instead of absolute values, the different resolutions of the models do not influence these results."*

The following sentence was added to the "Model Weighting" section in the methods L200-201: *"The weights of the models were used to create the multi-model weighted ensemble means for figures 1, 2, 4, S3, S7-S9, and S23-S25."*

To the section "Regional Analysis" L288-291 in the methods we added: "*These IPCC WGI regions were used to calculate the weighted multi-model ensemble regional means, for tables 1, S3 and S4. Furthermore, weights were applied to reflect the cell sizes, so that cells with larger land masses (around the equator) get higher weights than the cells with small land masses (higher latitudes around the poles)."*

* The introduction could be improve by better embedding in existing literature. Some papers discuss the larger changes to the more rare extremes (like Aalbers et al, Hodnebrog et al,) and I guess there are more. Also Li et al already performed an elaborate analysis of changes in precipitation extremes in CMIP6.

Thank you for directing us to more literature. We have rephrased our introduction L40-48 to include these works: *"The second type of extremes are the "rare" ones with multi-year or multi-decade return times, which are important for infrastructure design [10]. In hydrology these are typically estimated based on extreme value theory (using a historical time series of the same location), but other model-based (e.g. [26, 27]) or spatial pooling based approaches (e.g. [28, 29]) also exist to increase the time series length. There are fewer studies on the effect of global warming on such rare extremes [30, 31 , 32], or on the differences in future changes between "common" to "rare" extremes [15, 27], but the latter point to a possible larger relative increase of the rare extremes"*

We have also included references to some of these works to the results section L87-90: *"These regional patterns of increasing and decreasing precipitation extremes are similar to those of Li et al. [15], their figure 5. Furthermore, the areas with the highest increases and lowest decreases overlap with the areas with the most positive and negative scaling with temperature [46], their figure 4."*

And in the results L102-104: *"As well as based on a single model initial-condition large ensemble study over Western Europe [27], and on CMIP6 global climate models [15]."*

minor:

* L 23-33. Perhaps would be good to mention work on CMIP6 and precipitation extremes, such as in Li et al (2021). This could be at the expense of the discussion on small-scale/short duration extremes. I am not sure whether you can really say that daily are good enough represented in the CMIP models, considering the that in the tropic these are of convective origin. I'd rather say something like that despite these limitations, GCMs are still the only available tools to look at (rare) extremes at a global scale.

Rephrased L25-30 to: *"Global climate models (GCMs) are the only available tools to study future daily rainfall extremes on the global domain, but come with limitations. A large limitation is that GCMs do not resolve convective processes, which are important drivers of extreme precipitation [12, 13]. Recent research demonstrates that GCMs included in the Coupled Model Intercomparison Project Phase 6 (CMIP6) [14] have decent skill in modelling extreme rainfall in comparison to observations [15]."*

* L31 it's rather large-scale dynamical or synoptic scale dynamical. Also for small-scale events changes to the dynamics are very important, perhaps even more than for the large-scale events.

We removed the section on small-scale convective/large-scale dynamical precipitation because of the previous comment.

* L42. Extreme value theory are a tool to study these rare extremes, although in principle that can also be derived from single model initial condition large ensembles (SMILEs) or spatial pooling of results. Perhaps formulate this more generally.

Rephrased L40-44 to: "*The second type of extremes are the "rare" ones with multi-year or multi-decade return times, which are important for infrastructure design [10]. In hydrology these are typically estimated based on extreme value theory (using a historical time series of the same location), but other model-based (e.g. [26, 27]) or spatial pooling based approaches (e.g. [28, 29]) also exist to increase the time series length."*

* L97. Would be valid information to also have the change in the 1-year return level in this table

Table 1 shows $D_{100\text{-}1,x}$ (eq. 3), which is the difference in the change in 100-year return level ($C_{rel,T100,x}$, eq. 2), minus the change in the 1-year return level ($C_{rel,T1,x}$, eq. 2). For clarification, we have now also included the change in 1-year return levels ($C_{rel,T1,x}$, eq. 2) and change in 100-year return levels ($C_{rel,T100,x}$, eq. 2) for all SSP scenarios in Tables S3 and S4 in the supplement, and added links to these tables in the text.

Reviewer #2 (Remarks to the Author):

In this paper the authors aim to assess if rare extremes (return period 100 years) will relatively more than common extremes corresponding in return periods of 1 year. The question is relevant, but from a statistical point of view, given the assumption that warming will cause changes in precipitation regime, the answer is yes. Assuming a nonstationary distribution expressing this change then the relative change, in general, cannot be the same at T =1 year and T = 100 years, or for different T's in general (I provide details later on). Yet whether it's going to be larger or smaller change in one of the other case depends on the type of nonstationarity. Overall, the paper is well-written and well organized and deserves communication. Next I provide some specific comments that the authors might want to consider.

Many thanks to the reviewer for their time, insightful comments, and finding our study relevant, well-written and well organized, and worthy of communication. We respond to each of the points raised below.

42 annual maxima are still based on extreme value theory

In what is now L35 we referred to annual maxima as simply taking the largest value per year, without fitting an extreme value distribution. In what is now L40 we referred to extreme value theory, which could use annual maxima as input to fitting an extreme value distribution, but this is not a necessity.

139 just to be 100% sure by the historical late 20th century period 1971-2000 the authors refer to the simulated historical as mentioned in 52. I would add it again to avoid confusion.

Thank you for pointing this out, we have updated the text accordingly in L159: "The two time periods that are compared are (1) the *simulated* historical late 20th century period 1971-2000, …"

140-141 A general point is why include only the last 30 years of the 21th century. A policy/decision maker might be also interested in the next 30 years maybe even more than 70 years ahead.

Thank you for raising this point. The scientific point in our opinion is clearest when comparing just these two periods in time. Using our findings in practical products for policy/decision makers should indeed incorporate the near future as well, but this can be considered a goal of future (applied) research. Moreover, it would take a significant amount of time to analyse other 30 year time blocks and we have, therefore, not done this for this resubmission. However, we would be happy to do so if the editor thinks this would be of significant added value to our manuscript.

142-143 what is the rational of selecting the first of the available simulation of each model? Why not select the best simulation in terms of observations? I mean compare simulation with observed data and select the best performing.

We admit that just taking the first available simulation is somewhat arbitrary. However, going through an additional evaluation step would not add much to the manuscript as in fact we have shown all individual simulations in the supplementary information (Figures S12-S22) and found large similarities. Therefore, the choice for the first available simulation of each model for the averaging and final maps was pragmatic, but does not have any effect on our conclusions.

To clarify this in the manuscript, we have added the following to L164-L167: *"We arbitrarily selected the first available realisation of each GCM, however, the similarity between different realisations (Figs. S12-S22) led us to believe that this did not affect our main findings."*

146 some models have quite coarse resolution; I believe remapping at 0.25 x 0.25 could affects the results. At least the authors should discuss these points and mentioned potential issues that might emerge.

Thank you for pointing this out. We have added the following text to include this point in L169-174: "*The results remain mostly unaffected by the remapping, as the 0.25° × 0.25° grid is a higher resolution than the native grid of each of the models. Specifically, the remapped grid cells are 4 to 63 times smaller depending on the model resolution, and as we used nearest neighbour interpolation no spatial averaging is taking place. Moreover, as the main results of this study are relative values instead of absolute values, the different resolutions of the models do not influence these results.*"

Model Weighting. My understanding is that this is a popular weighting scheme that have been adopted and is based on the 5 points mentioned. How these points are really important for precipitation extremes? Wouldn't be more important to test if the model reproduces the actual observed distribution of extremes? Variability is important in precipitation but you can have same variability and different tails for example.

The reviewer is right, ideally we would test the models' ability to simulate the real extreme distribution. This is, however, challenging in the historical period due to the lack of long observational time series and impossible in the future. Therefore, we base our assessment of CMIP6 models on metrics that we deem important for a skillful representation of extreme events namely climatology and variability of temperature and precipitation. As an additional metric we include the temperature trend, which has recently been found to be skillful in constraining changes in mean precipitation (Shiogama et al. 2022).

In general, constraining future precipitation changes due to climate change remains challenging (see e.g., Pendergrass 2020 for a discussion of weakening or disappearing constraints proposed for CMIP5 in CMIP6). Here we use the performance weighting mainly to guard against models with large differences to the observations in one or more of the selected metrics as we consider them less trustworthy. In addition, we also use the independence weighting part to account for dependencies between the models that can arise due shared components or other overlaps in their source code. We have complemented the methods section L186-200 with additional information:
"For model performance we adapted the metrics used in [35] to the target of global precipitation change. In contrast to other important climate variables, most prominently future warming [35, 55, 56, 57], emergent constraints [58] for global precipitation changes have only recently been suggested [59, 60]. Here, our main aim is to reduce the influence of models which simulate variables considered important for the representation of precipitation very different from observations rather than applying a constraint that necessarily reduces model spread. Performance weights were, therefore, based on five metrics: 1) the temperature trend, which has been found to be an important constraint for temperature and precipitation changes alike [55, 59], 2) the temperature climatology, 3) the variability of temperature, 4) the precipitation climatology and 5) the variability of precipitation, all in the period 1979-2014. Models which perform badly in one or more of these metrics receive less weight in the calculation of multi-model statistics as we trust their projections of future precipitation change less. The strength of the weighting is established using a leave-one-out model-as-truth test [37, 61, 35] on the target of global mean precipitation change. The resulting weights for each model

are included in Table 2 and have a range comparable to recent studies, such as Brunner et al. [35] (their table S2 in the supplement). The weights of the models were used to create the multi-model weighted ensemble means for figures 1, 2, 4, S3, S7-S9, and S23-S25."

Furthermore, in our revised manuscript (and in response to your final comment) we do refer to a study by Li et al. (2021) where they did look at how different climate models included in CMIP6 simulate the 50-year return level in comparison to ERA5 reanalysis data. They found that the models included in CMIP6 do have reasonable skill in simulating precipitation extremes, which leads us to believe that the precipitation in the global climate models have similar tails to those in this reanalysis dataset.

183-185 well this could be the case only if the tail used in MEV matches reality. The Weibull tail shows in general good results, but other studies show also good results for power-type tails. So maybe add this assumption after the "is therefore able to estimate return periods higher than the period of record with reduced uncertainty" if the tail matches. The fact that you might get smoother patterns does not mean that you don't get consistent bias which you'd get as I mentioned if there's a mismatch in the asymptotic behavior.

Thank you for your comment. We have rephrased the text at L219 accordingly: "As opposed to traditional extreme value distributions, MEV uses all available data, and is, therefore, able to estimate return periods higher than the period of record with reduced uncertainty *if the tail of the true distribution matches*".

Statistical Analysis. One could approach this problem a bit differently or use another method to verify the results. The question of whether the rarest extremes will increase more than the common ones can be also replied if one investigates the nonstationary GEV or any other nonstationary distribution. For example if you use the historical simulation + the future simulation and fit a nonstationary gev then you get the answer to your question easily since you can estimate the relative change as 100(Q_GEV(a(t2),b(t2),c(t2),1-1/T)/ Q_GEV(a(t1),b(t1),c(t1),1-1/T)-1) for T =1 year and then for T =100 years and for any times t2 (e.g, 2100) and t1 (e.g., 2000). This will give you the answer. Of course different types of nonstationarities can be assumed here (e.g., changes in location parameter, or scales, or even shape (not advised) or combinations).

This proposed method is certainly viable, however, we have in our opinion already sufficiently verified our findings based on MEV, with an analysis based on GEV as well as directly model output for both the historical and future 30-year period. These results are shown in Figures S10 and S11 (we had some typos in the caption of figure S11 (previously S8), which is now corrected). In any case, we would be happy to see our results being further confirmed (or contrasted) by other statistical methods, which may come with other challenges (e.g. Li et al. (2019) demonstrated that 30-year periods are short to reliably fit non-stationary models), but leave this up to future research.

A final point of discussion which I might have missed is if there is any validation of the simulations with actual observations. I mean CMIP6 or climate models in general sometimes predict precipitation from another planet, what is the point of the analysis if the models don't predict well extreme precipitation? To clarify, I understand that you wish to investigate relative changes, yet since the projections have not been bias corrected then results shown might be misleading. For example results shown in Fig 2 and 3 or mentioned in the abstract and discussion can give a wrong impression. My understanding is that you have not bias corrected the simulations and have not compared with historical observations. Thus the analysis shows the relative change in the simulation of a model; if the model is bad then such numbers can be misleading. The D measure seems better in this sense.

It is true that we have not validated the simulated precipitation with observed precipitation, as this unfortunately is not so straight forward, and one has to consider aspects as which observed precipitation product that is completely homogeneous for the entire period of time, different grid-cell sizes for the different models and observed precipitation products, no future observations available, etc. We do refer to the work of Li et al. (2021) for the verification of 1-day precipitation estimates with a 50-year return level for a large CMIP6 multi-model ensemble with respect to the "observed" precipitation in the ERA5 reanalysis dataset. The years they are comparing are 1985-2014, so different from our study (1971-2000). Their figure 3(a, c, e) shows that the CMIP6 models capture the precipitation extremes reasonably well, predominantly over the extra tropics, but that there are biases in magnitude over the tropics and subtropics. These biases are that the climate models simulate precipitation extremes as being too much in the subtropical dry regions (northern Africa and eastern subtropical basins of the South Atlantic and South Pacific), and too little in the intertropical convergence zone. The correlation between the 50-year return levels for daily precipitation events is 0.88, so the simulated extremes are reasonably similar to those in ERA5.

Moreover, although we have not bias-corrected the precipitation, our weighting approach warrants to some extent for taking into account models that are far from reality (by giving them a lower weight). Our results (Figures 2 and 4) also show that in several regions there is disagreement between models. Before using our results in a practical situation for future extreme rainfall-intensity-frequency curves for particular locations, we agree that an even more careful selection of models based on regional observations and advanced bias-correction techniques is crucial. We have included the following to the introduction section L30-32: *"Yet, when interested in absolute magnitudes or specific locations, a careful selection of models based on observations or advanced bias correction approaches are necessary….."*

Furthermore, we have added the following text to the discussion L132-136: *"It should be noted that while the larger patterns of rarer extremes increasing relatively more is quite robust, there are also some regions with model disagreement. For such regions particularly, when compiling future extreme rainfall-intensity-frequency curves a more careful selection and weighting of climate models based on regional observations and advanced bias-correction techniques is advisable."*

Overall, I believe this is a decent paper and deserves communication.

Thank you.

References

Li, F. W. Zwiers, X. Zhang, and G. Li. (2019) How much information is required to well constrain local estimates of future precipitation extremes? *Earth's Future*, 7, 11–24, https://doi.org/10.1029/2018EF001001

Li, C., Zwiers, F., Zhang, X., Li, G., Sun, Y., & Wehner, M. (2021). Changes in annual extremes of daily temperature and precipitation in CMIP6 models. *Journal of Climate*, *34*(9), 3441-3460. https://doi.org/10.1175/JCLI-D-19-1013.1

Pendergrass, A. G. (2020). The Global-Mean Precipitation Response to CO2-Induced Warming in CMIP6 Models. *Geophysical Research Letters, 47*(17), 1–10.,https://doi.org/10.1029/2020GL089964

Shiogama, H., Watanabe, M., Kim, H., & Hirota, N. (2022). Emergent constraints on future precipitation changes. *Nature, 602*(7898), 612–616, https://doi.org/10.1038/s41586-021-04310-8

Thackeray, C. W., Hall, A., Norris, J., & Chen, D. (2022). Constraining the increased frequency of global precipitation extremes under warming. *Nature Climate Change, 12*(May), https://doi.org/10.1038/s41558-022-01329-1

11th Aug 22

Dear Dr van der Ent,

Your manuscript titled "Rarest rainfall events have greatest relative increase under climate change" has now been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment under the open access CC BY license (Creative Commons Attribution v4.0 International License).

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

- Please confirm that the increase of daily land rainfall extremes reported is in term of magnitude rather than frequency and ensure that whichever is correct is clearly stated in both the title and abstract.

- Please proofread your manuscript carefully, making sure that no grammatical errors or typos have remained in the manuscript.

- Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication under a CC BY license (Creative Commons Attribution 4.0 International License). This license allows maximum dissemination and re-use of open access materials and is preferred by many research funding bodies.

For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <a href="https://www.nature.com/commsenv/article-

processing-charges">https://www.nature.com/commsenv/article-processing-charges

At acceptance, you will be provided with instructions for completing this CC BY license on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:
[link redacted]
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Alienor Lavergne, PhD
Associate Editor
Communications Earth & Environment

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

I am very satisfied with the revisions, and fully support publication of this paper. I got two remarks left:

abstract 16-17. Would be good to state clearly that this is magnitude increase (not frequency, which is probably much more)

l 98. This is a very interesting observation, but would be good to mention it is actually dew point temperature in Ali et al (over land that could make a large difference)

Reviewer #2 (Remarks to the Author):

I have no further comments. Thank you for replying in detail and clarifying the points I raised.

EDITORIAL REQUESTS:
- Please confirm that the increase of daily land rainfall extremes reported is in term of magnitude rather than frequency and ensure that whichever is correct is clearly stated in both the title and abstract.
We confirm that the increase is in magnitude, and have adjusted the title to the suggested: “Rarest rainfall events will see the greatest relative increase in magnitude under future climate change”. We have clarified this throughout the manuscript.

- Please proofread your manuscript carefully, making sure that no grammatical errors or typos have remained in the manuscript.
We have proofread the manuscript and believe there are no more grammatical errors or typos.

- Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files.
We have done so, see the attached document.

REVIEWERS' COMMENTS:
Reviewer #1 (Remarks to the Author):

I am very satisfied with the revisions, and fully support publication of this paper. I got two remarks left:
Thank you for your kind words and time in improving our manuscript by providing two more constructive comments.

abstract 16-17. Would be good to state clearly that this is magnitude increase (not frequency, which is probably much more)
We agree it would be better to more clearly state that we look at magnitudes, and have changed abstract 16-17 to: “For the mitigation scenario SSP1-2.6 and the high emission scenario SSP5-8.5 **the magnitude of** daily land rainfall extremes will increase by 10.5 % and 28.2 %, respectively, for yearly events and by 13.5 % and 38.3 %, respectively, for 100-year events by the end of this century.”

l 98. This is a very interesting observation, but would be good to mention it is actually dew point temperature in Ali et al (over land that could make a large difference)
Many thanks for this comment. We have rephrased the text to: “Furthermore, the areas with the highest increases and lowest decreases overlap with the areas with the most positive and negative scaling with **dew point** temperature \cite{Ali2021}, their figure 4.”

Reviewer #2 (Remarks to the Author):

I have no further comments. Thank you for replying in detail and clarifying the points I raised.
Many thanks for your time spend reviewing and improving our manuscript.