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

20th Apr 21

Dear Dr Brunner,

Your manuscript titled "Reconciling the extreme precipitation-flood paradox in a warming climate" has now been seen by 2 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that fully addresses these serious concerns.

We hope you will find the reviewers' comments useful as you decide how to proceed. Should additional work allow you to
• address these criticisms (that is, either to incorporate the suggestions or provide a compelling argument why the point made by the reviewer is not valid, or relevant to the editorial threshold as outlined below)
AND
• meet our editorial thresholds as outlined below,

then we would be happy to look at a substantially revised manuscript.

In the following, we list our main editorial requirements.

****
****Editorial threshold 1: Make a compelling case for threshold behaviour in the flood response to increasing rainfall, for the region you study (fully addressing reviewer 2's point (1)), and appropriately discuss potential applicability of your findings to other regions based on the existing literature.

****Editorial threshold 2: Explain your methods to a level of detail that will allow an independent researcher to reproduce your results.

****

However, please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

If the revision process takes significantly longer than three months, we will 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.

We are committed to providing a fair and constructive peer-review process. Please do not 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 any 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 **

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further. Thank you for the opportunity to review your work.

Best regards,

Heike Langenberg, PhD

Chief Editor
Communications Earth and Environment

On Twitter: @CommsEarth

EDITORIAL POLICIES AND FORMAT

If you decide to resubmit your paper, please ensure that your manuscript complies with our editorial policies and complete and upload the checklist below as a Related Manuscript file type with the revised article:

Editorial Policy Policy requirements

For your information, you can find some guidance regarding format requirements summarized on the following checklist:(https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf) and formatting guide (https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf).

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

1. In section 1.1, readers will need more details of the hydro-SMILE method.
2. Line 86 (Figure 1): It is necessary to describe how the relative changes are estimated. What is the base period? Is it estimated for future ensembles against historical record. This kind of information is not provided anywhere.
3. Line 87-95: Without any description of hydro-SMILE, it will be difficult for readers to understand how ensembles are generated from different climate and hydrological models. Nowhere it is mentioned how many ensembles are generated. Is it possible that section 4 (Methods) is given before Section 2(Results)?
4. In Figure 5, what are the parameters used to decide median and moderate flood?
5. Line 218: Threshold of what? This has not been mentioned anywhere.
6. Nowhere in the manuscript it is mentioned how relative changes are linked with P-Q response.
7. No description of model calibration is found.
8. What are the resolution of hydrological and meteorological models? How changes in resolution may impact the outcome of this study?
9. Please list major limitations of this study.
10. How the hydrological model uncertainties are accounted in this study? Do you think if a different hydrological model is used, we would have got the similar result?
11. How return periods are estimated for both precipitation and floods?

Reviewer #2 (Remarks to the Author):

I was a reviewer of this manuscript when it was submitted to Nature Climate Change. I will reiterate that the manuscript addresses a critical question that has perplexed hydrologists - the apparent lack of spatially-consistent flood increases when precipitation (particularly heavy downpours; see USGCRP, 2018) are increasing and climate models are generally in consensus about precipitation trends. There have been a number of papers looking to address this question (as the manuscript properly cites); however, no definitive, quantitative and robust explanation has solved this 'paradox,' as the authors nicely name this issue.

Therefore, a quantitative, authoritative, and robust result that would bring clarity to this topic would be a welcome and highly impactful contribution. In my previous review, I noted two major concerns with the manuscript: (1) the use of model coefficients estimated from multiple linear regression to evaluate the effects of antecedent conditions on floods, and (2) the use of only 78 catchments in hydrological Bavaria, to derive the conclusions.

My overall impression of the major conclusions is that larger flood events reach a tipping point where rainfall becomes the dominant process and swamps the effects of land surface processes, such as land cover (lines 217-220). This tipping point - termed 'extremeness' in the text - is this tipping point and is associated with a particular return period.

However, I am not sure how novel this finding is without more quantitative and robust findings. Most of the extreme flooding the US - for example - has come from periods of long-duration rain events (for example, the flooding of Houston, Texas, USA during Hurricane Harvey: https://www.climate.gov/news-features/event-tracker/reviewing-hurricane-harveys-catastrophic-rain-and-flooding) or short duration but intense storms (for example, the Eillicott City, Maryland, USA flood: https://www.bloomberg.com/news/articles/2019-05-24/a-historic-river-town-confronts-a-flooded-future). The conclusion that at some return interval, floods are dominated by precipitation does not present as novel.

It would be more useful to present an approach that provides a way to determine the precise (as possible) extremeness threshold for any given catchment and show that this approach is valid for many different types of catchments at a continental or global scale. From the results presented, it is not clear how one would determine the extremeness threshold for a catchment (the study appears to only show that one may exist) and there is only limited testing of this idea to 78 catchments in hydrological Bavaria.

There are aspects of this manuscript that are certainly worthy of publication in the scientific literature but I believe these limitations make it difficult to recommend the manuscript for further consideration in Communications Earth & Environment.

As I had recommended in my other comments, I would encourage the authors to push for a more quantitative and robust analysis by: (1) applying more rigorous regression methods, including panel methods, and (2) expanding the dataset to a continental or global scale. I offer more detailed comments on these points as well other comments below.

USGCRP, 2018: Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II [Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 1515 pp. doi: 10.7930/NCA4.2018.

Other Major Comments:

(1) It appears from lines 137-159, Section 1.6, and Figure 5 that conclusions are drawn from the use of the multiple linear regression (MLR) coefficients. When MLR is used in a prediction mode, the values of the coefficients are those that provide the best fit to the data; however, for the purposes of this study, the estimates of the coefficients must be reliable as measures of the the sensitivity of floods to precipitation and land-surface properties. The estimated coefficients from MLR are subject to a number of issues such as multicolineary and omitted variable bias (see Steinschneider et al., 2012 and Blum et al. 2020 for a discussion on this) and, therefore, are not the best representation of sensitivity of floods to the independent variables in the regression. While the authors do report that multicollinearity is not necessarily an issue, the larger issue of omitted variable bias and its effect on the regression coefficients is not handled.

Furthermore, if MLR were still to be used, it is not clear that the regression coefficients were checked to determine if they were significantly different from zero (meaning they actually provide some predictive relationship)? This would require that the residuals of the MLR were normally-distributed, heteroscedastic and independent. In cases where the regression coefficients were not significant, I am not sure if future changes in those drivers should even be reported. Then to compare the regression-estimated coefficients of past drivers with the regression-estimated coefficients of future drivers appears even more uncertain. I believe it is possible in the MLR process to at least have the uncertainties in the regression coefficients reported (assuming again that all of the assumptions of the residuals have been satisfied).

A panel regression approach could be applied to obtain better estimates of the coefficients and would allow for the inclusion of time as another variable in the regression. This approach would lead to more confidence in the regression coefficient estimates in terms of their causal attribution and isolate the effects of the confounding variables (precipitation versus land-surface variables).

If panel regression methods were applied, these methods have the advantage of not being concerned with the statistical significance of the coefficients, and omitted variable bias would not be an issue.

Blum, A.G., P.J. Ferraro, S.A. Archfield, and K.R. Ryberg, 2020, Causal effect of imperviousness on annual flood magnitude for the United States, Geophysical Research Letters, https://doi.org/10.1029/2019GL086480.

Steinschneider, S., Y.-C. E. Yang, and C. Brown (2013), Panel regression techniques for identifying impacts of anthropogenic landscape change on hydrologic response, Water Resour. Res., 49, 7874–7886, doi:10.1002/2013WR013818.

(2) The use of 78 catchments in Bavaria seems quite limited despite explaining that there is considerable variability. The argument that these catchments represent ‘diverse catchment characteristics’ is made a number of times, for example on lines 182-183 and lines 223-236, with an attempt to apply broader and unsupported conclusions that these “results suggest that many regions around the globe” would exhibit similar threshold behavior.

Unless you are able to show a comparison of the range of characteristics for these 78 catchments to the range of catchment catchments across a continental or global scale, it is not convincing that these 78 catchments represent sufficient variability to generalize the conclusions. For example, elevation or season may not be relevant for other areas in explaining threshold behavior (line 185).

Another way in which the results may be limited by the dataset is in the regression results. In regression equations for different flood recurrence intervals across the United States, there are regions where precipitation statistics are not included in the regression estimates because they are not statistically significant in the estimation of floods. I wonder if other areas of the globe were chosen, you may have situations where precipitation would not be a factor in regression-based flood estimates.

In my opinion this is a serious limitation of the assessment and a broader set of results would be needed to support robust conclusions, consistent with the scope of Communications Earth & Environment.

(3) To validate your conclusions, particularly if a larger study area with more gauges was used, could one find locations where there has already been a large change in flood magnitudes and compare the characteristics of those gauges with these findings?

(4) It seems that the term ‘extremeness’ is another term for return interval (line 68). I think that it would be more intuitive to simply use the return interval rather than invent a new (and somewhat awkward) word. It would be simple to say that there exists a “return interval threshold” or a “threshold based on the return interval.”

Responses to Reviewers

Reviewer #1:

1. In section 1.1, readers will need more details of the hydro-SMILE method.
Reply: *Thank you very much for highlighting the need to provide more details on the hydro-SMILE method already in the main part of the manuscript instead of just mentioning them in the Methods section (4.2) at the end of the manuscript. We added the following overview on the hydro-SMILE approach to Section 1.1. and indicate that further, more detailed information can be found in Section 4.2, which we renamed from 'Data' to 'Hydro-SMILE' in order to indicate that further information on the hydro-SMILE are provided in this section: 'The hydro-SMILE consists of hydrological simulations obtained by driving a hydrological model with climate simulations from a single model initial-condition large ensemble (SMILE) climate model. The underlying model simulations were originally generated by [83] as part of the ClimEx project [38]. The hydro-SMILE simulations consist of daily streamflow (mm/day), soil-water-equivalents (SWE, mm), and soil moisture (%) – all of which were obtained by driving the hydrological model WaSiM-ETH [64] with a 50-member ensemble of high-resolution climate input (spatial: 500 x 500 m^2, temporal: 3 h) (for further information on the hydro-SMILE see Section 4.2).'*
Modification: *p. 3, l. 72-79*

2. Line 86 (Figure 1): It is necessary to describe how the relative changes are estimated. What is the base period? Is it estimated for future ensembles against historical record. This kind of information is not provided anywhere.
Reply: *We agree that a description of how relative changes are computed should already be provided in the main part of the text and not just in the Methods section (Section 4.5 l. 380-388). We therefore added the following specification to figure captions 1-5: 'Relative changes are computed by comparing event characteristics of a future period (2060-2099) to characteristics of a historical period (1961-2000).'*
Modification: *figure captions 1-5*

3. Line 87-95: Without any description of hydro-SMILE, it will be difficult for readers to understand how ensembles are generated from different climate and hydrological models. Nowhere it is mentioned how many ensembles are generated. Is it possible that section 4 (Methods) is given before Section 2 (Results)?
Reply: *CEE's 'style and formatting guide' (https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf) asks for a manuscript structured into Abstract, Introduction, Results and Discussion, and Methods. We followed this guide and provide the Methods section at the end of the manuscript. We agree, however, that it might be beneficial for the reader to get a short introduction to the hydro-SMILE before moving to the results section. Therefore, we expanded Section 1.1., which is part of the Introduction, by providing a short description of the hydro-SMILE including information about the climate ensemble, the hydrological model, and the ensemble size (50 members) (also see our response to point 1). This should well prepare the reader for the Results section. We then explicitly refer the reader to Section 4.2 for more detail on the hydro-SMILE.*

Modification: *p. 3, l. 72-79*

4. In Figure 5, what are the parameters used to decide median and moderate flood?
Reply: *Thank you for highlighting the need for clarification. We specified that by 'moderate' flood, we mean 'median over all events identified in a catchment'.*
Modification: *p. 11, caption Figure 6*

5. Line 218: Threshold of what? This has not been mentioned anywhere.
Reply: *We specified that by 'threshold', we refer to 'an extremeness or return interval threshold.'*
Modification: *p. 11, l. 240*

6. Nowhere in the manuscript it is mentioned how relative changes are linked with P-Q response.
Reply: *Section 4.5. provides detailed information on how relative changes in P and Q magnitudes and their dependence were computed, which might not have been made clear in the main section of the text. We therefore added a short description on how we computed relative changes to the first paragraph of the results section where we also added a reference to Section 4.5: 'We then empirically compute P and Q magnitudes for different levels of extremeness, i.e. mean events and progressively more extreme events with 10, 20, 50, 100, and 200 year return intervals and finally derive future relative changes in extreme event magnitudes by comparing magnitudes for a future period (2060-2099) with magnitudes of a historic period (1961-2000) (see Methods 4.5).' We furthermore specified in captions 1-5 that 'Relative changes are computed by comparing event characteristics of a future period (2060-2099) to characteristics of a historical period (1961-2000).'*
Modification: *p. 4, l. 104-108 and figure captions 1-5*

7. No description of model calibration is found.
Reply: *We substantially expanded the paragraph about model calibration in Section 4.2, which now describes the spatial data sets used as model input and provides references to each of these data sets, specifies which datasets were used for calibration at which resolution, specifies the different components of the objective function used and their respective weights, and refers to the optimization algorithm used for parameter calibration: 'The model was set up for 98 catchments in Hydrological Bavaria by [83] using spatial information on elevation, slope and exposition derived from a digital elevation model for Europe (EU-DEM; [28]), land-use derived from the CORINE land cover data set [27], soil characteristics derived from the European soil database (ESDB v2.0; [29]), and hydro-geology (hydraulic conductivity) derived from the Bavarian hydrogeology map [4] and the international hydrogeological map of Europe (IHME1500 v1.1.; [16]) to define global model parameters (i.e. parameters applied to all 98 catchments) describing evapotranspiration rates, infiltration rates, groundwater fluxes, snowmelt, and glacier dynamics and by calibrating four parameters, i.e. those related to recession and direct flow. These local parameters were calibrated for the period 2004–2010 using the dynamically dimensioned search algorithm [71] on observed 3-hourly discharge of the 98 catchments provided by the Bavarian Environment Agency (Bayerisches Landesamt für*

Umwelt - LfU; [5]) and sub-daily observed interpolated meteorological input (i.e. precipitation, temperature, relative humidity, incoming shortwave radiation, and wind speed).
The meteorological Sub-Daily Climatological REFerence dataset (SDCLIREF) created in the ClimEx-project is based on a combination of hourly and disaggregated daily station data. To obtain the disaggregated daily station data, the method of fragments [80] was used to extend the sub-daily record to 1981–2010 and to densify the station network. The station data were then interpolated to a 500 x 500 m^2 grid using a combination of multiple linear regression, considering elevation, exposition, latitude, and longitude, and inverse distance weighting [similar to 62]. The dynamically dimensioned search algorithm used a multi-objective function targeted at optimizing flood characteristics composed of the Nash–Sutcliffe efficiency (ENS; [54])) and the Kling–Gupta efficiency (E_{KG}; [30]), which both focus on high flows [48], the log(E_{NS}), which emphasizes low flows, and the root-mean-squared error to standard deviation ratio (R_{SR}), which quantifies volume errors. The overall objective function assignes a lot of weight to the metrics E_{NS} and E_{KG} because our study focuses on flood events:
$M= 0.5×(1−E_{NS}) + 0.25×(1−E_{KG}) + 0.15×(1−\log(E_{NS})) + 0.1×R_{SR}$.'
Modification: *p. 14, l.358-370*

8. What are the resolution of hydrological and meteorological models? How changes in resolution may impact the outcome of this study?
Reply: *Thank you for pointing out the need to better address the resolution of the meteorological and hydrological data/models. We specified in the Introduction that the hydrological model is driven with meteorological input at a spatial scale of 500 x 500 m^2 and at a temporal resolution of 3h. We furthermore specified at the beginning of the results section that 'we use a hydro-SMILE consisting of a 50-member ensemble of 3-hourly precipitation and streamflow time series for Hydrological Bavaria, which we aggregated to daily resolution.' We also added this aggregation step to the Methods section: 'For the subsequent analyses of extreme precipitation and flood events, the 3-hourly meteorological and streamflow time series were aggregated to a daily scale and averaged over each catchment.'*
The current RCM setting uses parameterization for convective precipitation (which applies to all RCMs coarser than 3-4km and of course to GCMs). New convection permitting models might result in different precipitation changes over the study region considered. However, further increasing spatial resolution can introduce additional noise whose accounting for would require an ensemble of CPM simulations.
Modification: *p. 4, l. 97; p. 15, l. 393-394*

9. Please list major limitations of this study.
Reply: *We highlight model uncertainty as one limitation of our study in Section 1.1 (from l. 81): 'We acknowledge that the hydro-SMILE modeling chain is affected by uncertainties introduced through both the underlying climate and hydrological models. Climate model uncertainties include those relating to precipitation process-representation, downscaling, and bias-correction procedures, hydrological model uncertainties comprise model and parameter uncertainties. These latter uncertainties may be particularly relevant for the very extreme events under consideration in the present study because model calibration and evaluation rely upon observed events - and (as previously noted) modern observational records simply don't exist for events of*

the extreme magnitudes considered here. However, we point out that this particular element of the overall uncertainty is essentially irreducible, and will likely remain so until the length of the observed record increases substantially some decades in the future.'

We furthermore emphasize that our study was performed on a data set of catchments representative of the temperate climate zone and additional similar analyses will be required to confirm the threshold behavior in other climate zones: 'While these regime types can be considered representative of the temperate climate zone with similar runoff regimes (pluvial to nival), our catchment selection does not cover other climate zones such as cold-climates, semi-arid to arid regions, and the tropics.'
Modification: *p. 13, l. 343-345*

10. How the hydrological model uncertainties are accounted in this study? Do you think if a different hydrological model is used, we would have got the similar result?
Reply: *Previous hydrological climate impacts focusing on uncertainty decomposition have shown that climate model and natural variability dominate total uncertainty and that hydrological model uncertainties are comparably small in most catchments (Addor et al. 2014; Vetter et al. 2017; Hattermann et al. 2018). Furthermore, Prein and Pendergrass (2019) as well as Lehner et al (2020) have shown that internal climate variability is a dominant source of uncertainty in climate impact assessments on regional and local scales. This inherent uncertainty of the climate system is irreducible and can only sufficiently be addressed by using SMILEs (Deser et al., 2012). While our analysis considers this important sources of uncertainty (natural variability) by using a SIMLE, it does neither consider climate model uncertainty as it applies one regional climate model to dynamically downscale the global climate model ensemble CanESM2-LE nor hydrological model uncertainty as we only use one hydrological model. It would be interesting to rerun the analysis with different combinations of SMILEs from different GCMs and hydrological models. However, such an analysis would require another careful model calibration process as well as substantial computational and data storage capabilities. We are confident that using a different GCM or hydrological model would not change the main conclusions of our analysis.*

11. How return periods are estimated for both precipitation and floods?
Reply: *Thank you for pointing out this omission. We specified in both the Results and Methods sections that the return periods were estimated empirically by pooling events identified in the 50 ensemble members: 'We then empirically compute P and Q magnitudes for different levels of extremeness, i.e. mean events and progressively more extreme events with 10, 20, 50, 100, and 200 year return intervals, by pooling events extracted from the 50 ensemble members.' 'Sample quantiles are computed for probabilities corresponding to different return periods T using: p=1-(mu/T), where mu is the mean inter-arrival time between events.'*
Modification: *p. 4, l.104-106; p. 16, l. 447-448*

Reviewer #2:

I was a reviewer of this manuscript when it was submitted to Nature Climate Change. I will reiterate that the manuscript addresses a critical question that has perplexed hydrologists - the apparent lack of spatially-consistent flood increases when precipitation (particularly heavy downpours; see USGCRP, 2018) are increasing and climate models are generally in consensus about precipitation trends. There have been a number of papers looking to address this question (as the manuscript properly cites); however, no definitive, quantitative and robust explanation has solved this 'paradox,' as the authors nicely name this issue.

Therefore, a quantitative, authoritative, and robust result that would bring clarity to this topic would be a welcome and highly impactful contribution. In my previous review, I noted two major concerns with the manuscript: (1) the use of model coefficients estimated from multiple linear regression to evaluate the effects of antecedent conditions on floods, and (2) the use of only 78 catchments in hydrological Bavaria, to derive the conclusions.

My overall impression of the major conclusions is that larger flood events reach a tipping point where rainfall becomes the dominant process and swamps the effects of land surface processes, such as land cover (lines 217-220). This tipping point - termed 'extremeness' in the text - is this tipping point and is associated with a particular return period.

However, I am not sure how novel this finding is without more quantitative and robust findings. Most of the extreme flooding the US - for example - has come from periods of long-duration rain events (for example, the flooding of Houston, Texas, USA during Hurricane Harvey: https://www.climate.gov/news-features/event-tracker/reviewing-hurricane-harveys-catastrophic-rain-and-flooding) or short duration but intense storms (for example, the Eillicott City, Maryland, USA flood: https://www.bloomberg.com/news/articles/2019-05-24/a-historic-river-town-confronts-a-flooded-future). The conclusion that at some return interval, floods are dominated by precipitation does not present as novel.

It would be more useful to present an approach that provides a way to determine the precise (as possible) extremeness threshold for any given catchment and show that this approach is valid for many different types of catchments at a continental or global scale. From the results presented, it is not clear how one would determine the extremeness threshold for a catchment (the study appears to only show that one may exist) and there is only limited testing of this idea to 78 catchments in hydrological Bavaria.

There are aspects of this manuscript that are certainly worthy of publication in the scientific literature but I believe these limitations make it difficult to recommend the manuscript for further consideration in Communications Earth & Environment.

As I had recommended in my other comments, I would encourage the authors to push for a more quantitative and robust analysis by: (1) applying more rigorous regression methods, including panel methods, and (2) expanding the dataset to a continental or global scale. I offer more detailed comments on these points as well other comments below.

Reply: *Thank you very much for your thorough assessment of both the present version and a previous version of our manuscript! Thank you also for highlighting the need to (1) present an approach that provides a way to determine the precise extremeness threshold for any given catchment, (2) improve the evaluation of the role of antecedent conditions on flood magnitudes by applying suitable regression approaches, and (3) discuss the results in a broader geographical context. We address these three main points in the substantially expanded version of our manuscript.*

To address point (1), we have added a new figure (Figure 2) to the manuscript showing future changes in flood magnitudes per return interval for each of the 78 catchments. This new figure highlights that the extremeness threshold above which precipitation increases result in discharge increases is catchment specific. This catchment specific information complements the existing figures, which summarize the threshold behavior by showing median changes in flood magnitudes/frequencies/dependencies over all catchments. We think that this additional figure demonstrates that our approach allows us to determine the extremeness threshold for any specific catchment independent of where it is located.

We would also like to highlight that the finding of an extremeness threshold above which increases in precipitation result in increases in flooding does not rely on the regression analysis, which is only applied in the second part of the manuscript to analyze the role of antecedent conditions in current and future flood response.

To address point (2), we show that the regression analysis applied is an appropriate method to study the role of different hydro-meteorological drivers for current and future flood magnitudes because it is only weakly affected by multicollinearity and unaffected by omitted variable bias.

Point 3 is the most difficult one to address because the hydro-SMILE presented here is to our knowledge the only existing hydro-SMILE for any region on the globe that has been generated with a physically based model at very high spatial and temporal resolution. We discuss in the Discussion Section how our analysis could be expanded to a global scale by calibrating a global hydrological model for flood events, though this would require an extremely large amount of computational resources. We also considerably expand our discussion and literature review regarding how our results derived using a sample of 78 catchments in Bavaria may (or may not) be generalized to other regions of the world with different hydro-climates. Please find our detailed responses to these major comments below.

USGCRP, 2018: Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II [Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 1515 pp. doi: 10.7930/NCA4.2018.

Other Major Comments:

(1) It appears from lines 137-159, Section 1.6, and Figure 5 that conclusions are drawn from the

use of the multiple linear regression (MLR) coefficients. When MLR is used in a prediction mode, the values of the coefficients are those that provide the best fit to the data; however, for the purposes of this study, the estimates of the coefficients must be reliable as measures of the the sensitivity of floods to precipitation and land-surface properties. The estimated coefficients from MLR are subject to a number of issues such as multicolineary and omitted variable bias (see Steinschneider et al., 2012 and Blum et al. 2020 for a discussion on this) and, therefore, are not the best representation of sensitivity of floods to the independent variables in the regression. While the authors do report that multicollinearity is not necessarily an issue, the larger issue of omitted variable bias and its effect on the regression coefficients is not handled.

Reply: *Thank you very much for pointing out the need to better demonstrate the appropriateness of multiple linear regression to study the role of potential hydro-meteorological drivers in determining flood response of moderate and extreme extremes. We totally agree that multiple linear regression (MLR) may be subject to multicollinearity and/or omitted variable bias if explanatory variables are not chosen carefully. We have conducted some additional analysis to check our MLR analysis for both multicollinearity and omitted variable bias and in the following demonstrate that our analysis is affected by neither of the two phenomena.*

To test for multicollinearity, we computed the variable inflation factor for the 4 explanatory variables in the 8 different models fitted for high and low elevation catchments in summer and winter for historical and future conditions. The summary statistics over all 32 VIF values are shown in the table below:

Minimum	*1st quartile*	*Median*	*Mean*	*3rd quartile*	*Maximum*
1.068	*2.016*	*3.154*	*3.483*	*4.183*	*9.187*

None of the explanatory variables shows a VIF value > 10 and hardly any explanatory variables shows VIF values > 4 in any of the model setups used to generate Figure 5 (now 6) in the manuscript, which indicates that multicollinearity is not an issue in our model setup. We add this information to the manuscript: 'VIF does not exceed 10 for any pair and only exceeds 4 for very few pairs.'

Modification: *p. 10, l. 195*

To test for variable omission bias, we compute the correlation between model residuals and each of the explanatory variables (see Figure 1). In the presence of variable omission bias, explanatory variables and model residuals would be correlated.

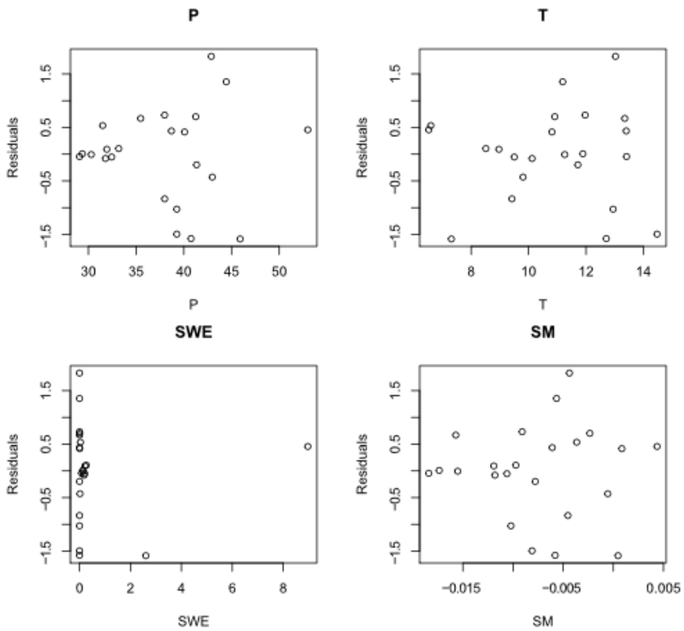

Figure 1: Scatterplots for model residuals of one of the 8 models and the four explanatory variables precipitation (P), temperature (T), snow-water-equivalents (SWE), and soil moisture (SM).

The correlation summary statistics over the 8 different model setups used are given in the following table:

Minimum	*1st quartile*	*Median*	*Mean*	*3rd quartile*	*Maximum*
-5.840e-17	*-4.422e-17*	*-2.063e-17*	*5.781e-18*	*2.704e-17*	*1.162e-16*

These results show that residuals are independent of the value of the explanatory variable. We can therefore assume that our analysis includes the most relevant predictors and is unaffected by variable omission bias. Because neither multicollinearity nor variable omission bias seem to be an issue in our case, we think that MLR is an appropriate method to study the relationship of flood magnitudes and the four hydro-meteorological variables precipitation, temperature, snow-water-equivalent, and soil moisture.

Furthermore, if MLR were still to be used, it is not clear that the regression coefficients were checked to determine if they were significantly different from zero (meaning they actually provide some predictive relationship)? This would require that the residuals of the MLR were normally-distributed, heteroscedastic and independent. In cases where the regression coefficients were not significant, I am not sure if future changes in those drivers should even be reported. Then to compare the regression-estimated coefficients of past drivers with the regression-estimated coefficients of future drivers appears even more uncertain. I believe it is possible in the MLR process to at least have the uncertainties in the regression coefficients reported (assuming again that all of the assumptions of the residuals have been satisfied).
Reply: *We agree that it is important to report the p-values associated with the individual regression coefficients. While we tested for significance in the previous version of the manuscript, we only reported significant vs. non-significant coefficients by hatching non-*

significant coefficients in Figure 5 (now Figure 6). This might not have been specific enough. The new version of figure 6 includes the actual p-values for each of the explanatory variables and models in addition to the hatching, which still indicates explanatory variables with p-values > 0.05.
Modification: *Figure 6*

A panel regression approach could be applied to obtain better estimates of the coefficients and would allow for the inclusion of time as another variable in the regression. This approach would lead to more confidence in the regression coefficient estimates in terms of their causal attribution and isolate the effects of the confounding variables (precipitation versus land-surface variables).
If panel regression methods were applied, these methods have the advantage of not being concerned with the statistical significance of the coefficients, and omitted variable bias would not be an issue.
Reply: *Thank you for suggesting an alternative regression approach to assess the role of hydro-meteorological drivers for flood magnitude. We agree that panel regression would be a suitable alternative to multiple linear regression if omitted variables were a problem. However, we have shown above (Figure 1) that such variable omission bias does not seem to be an issue for our analysis. We would therefore like to stick to our initial modeling decision of multiple linear regression, which represents the simplest and most parsimonious model possible. Furthermore, we think that even in a panel regression setting, certain of our explanatory variables may show insignificant coefficients.*

Blum, A.G., P.J. Ferraro, S.A. Archfield, and K.R. Ryberg, 2020, Causal effect of imperviousness on annual flood magnitude for the United States, Geophysical Research Letters, https://doi.org/10.1029/2019GL086480.

Steinschneider, S., Y.-C. E. Yang, and C. Brown (2013), Panel regression techniques for identifying impacts of anthropogenic landscape change on hydrologic response, Water Resour. Res., 49, 7874–7886, doi:10.1002/2013WR013818.

(2) The use of 78 catchments in Bavaria seems quite limited despite explaining that there is considerable variability. The argument that these catchments represent ‘diverse catchment characteristics’ is made a number of times, for example on lines 182-183 and lines 223-236, with an attempt to apply broader and unsupported conclusions that these “results suggest that many regions around the globe” would exhibit similar threshold behavior.

Unless you are able to show a comparison of the range of characteristics for these 78 catchments to the range of catchment catchments across a continental or global scale, it is not convincing that these 78 catchments represent sufficient variability to generalize the conclusions. For example, elevation or season may not be relevant for other areas in explaining threshold behavior (line 185).

Another way in which the results may be limited by the dataset is in the regression results. In regression equations for different flood recurrence intervals across the United States, there are regions where precipitation statistics are not included in the regression estimates because they are not statistically significant in the estimation of floods. I wonder if other areas of the globe were chosen, you may have situations where precipitation would not be a factor in regression-based flood estimates.

In my opinion this is a serious limitation of the assessment and a broader set of results would be needed to support robust conclusions, consistent with the scope of Communications Earth & Environment.

Reply: *Thank you for pointing out the need to make more nuanced statements about the generalizability of our findings. We agree that it would be desirable to expand our analysis to a global setting where all climate zones are represented. However, as we discuss in our discussion section, this would require a global hydro-SMILE (from l. 285): 'Creating such a global hydro-SMILE for flood analyses requires the combination of a globally downscaled and bias-corrected atmospheric SMILE with a global hydrological model specifically calibrated for flood peaks. Satisfactory calibration for far-from-mean state conditions is challenging using calibration metrics commonly used for large-scale model calibration (Brunner et al. 2021) and data storage and computational costs are high at a global scale when a large spatial domain is combined with a large ensemble size. In addition, global-scale models may not as accurately represent complex land-surface processes as smaller-scale models and appropriate reference datasets for meteorology, soils, and hydrogeology are harder to obtain. Creating such a global hydro-SMILE therefore remains a considerable research effort, but one of substantial importance in a warming climate.'*

We therefore rely our analysis on an existing hydro-SMILE, which represents catchments in the temperate climate zone with hydrological regimes ranging from nival to pluvial. We substantially expanded our Discussion by discussing the generalizability of our results to other climatic and land use contexts, as well as offering an expanded discussion of existing literature that offers some observational evidence for the existence of such a threshold in different climate regimes:

'In this work, we demonstrate for hydrological Bavaria that there is an extremeness or return interval threshold, which varies by catchment, season, and event type, above which extreme-precipitation increases outweigh the soil-drying effects of warming temperatures. This result suggests that in other regions around the globe with similar hydro-climates, i.e. temperate climates with pluvial or nival flow regimes, flood risk in a warming climate may also exhibit divergent changes above and below some locally-defined extremeness or return interval threshold. We further find that the hydrologic response to extreme precipitation varies predictably as function of event magnitude in a warming climate, with streamflow responses becoming increasingly positive even in the few study catchments which do not exhibit distinct threshold behavior. This, when viewed in the context of prior research, may offer evidence for broader geographic generalizability of our findings. We find that increases in precipitation yield larger and more consistent increases in flood magnitude for more extreme versus more moderate events—which is supported by previous observational studies showing only weak

dependence between extreme precipitation and moderate flood occurrence in the United States [24], stronger increasing flood trends for extreme than moderate floods in Central Europe [6], and trends in extreme discharge that only align with trends in floods for the rarest events in Australian catchments [79]. Thus, there does appear to be a growing body of real-world evidence suggestive of the existence of a precipitation-flood response threshold across a wider range of hydroclimatic and hydrologic regimes than explicitly considered in the present study. The complex influences of elevation, season, and event type upon the return interval threshold suggests that the location of this critical cross-over point may vary somewhat widely across regions of the world with varying topography and background climate. Substantial modulation of this threshold would likely occur depending on climatic factors such as aridity and the local relevance of snowmelt, catchment size, and land use and management. Consider, for example, a semi-arid or subtropical regime (as opposed to the moist mid-latitude regime that characterizes the catchments in the present study). In such a location, the return interval threshold might be higher due to drier antecedent soil conditions – a temperature-related phenomenon we also see when comparing seasonally-varying summer with winter thresholds (Figures 3, 4). The existence of a high return interval threshold in drier Mediterranean regions is supported by observation-based studies that have demonstrated a stronger relationship between precipitation and discharge for larger versus smaller flood events in Spain [12], and have shown decreases in the occurrence of moderate floods in southern Europe [8, 6]. In contrast, if we consider cold high-latitude regions and/or high altitude regions with a snow-dominant precipitation regime, the return interval threshold might be expected to be much lower. Indeed, this relationship is apparent from our threshold analysis for snow-influenced events in high-elevation regions (Figures 3, 4). Additionally, and as suggested by our results (Figure A.3) the return interval threshold may also be modulated by catchment area (generally increasing with catchment size). For larger river basins than the ones included in our Bavarian selection, this finding would imply higher return level thresholds than 20-50 years. Furthermore, direct human influence on streamflow such as dynamic reservoir operations and/or flood management interventions might lead to higher return interval thresholds because smaller floods can be buffered by temporary water storage [77]. In contrast, urbanized catchments (characterized by a high fraction of water-impervious surfaces) might have lower return interval thresholds than catchments with unsealed surfaces because of a more direct relationship between extreme precipitation and flood response [76, 10]. Such a return interval threshold might even vary from year to year in a single location—occurring at a higher level of event extremeness during drought versus pluvial periods.'
Modification: *p. 11 and 12, l. 240-282*

(3) To validate your conclusions, particularly if a larger study area with more gauges was used, could one find locations where there has already been a large change in flood magnitudes and compare the characteristics of those gauges with these findings?
Reply: *We agree that it would be great if long observed time series to study events with 100 or 200 year return periods were available. Unfortunately, most countries have few time series that extend beyond 30, 50 and 100 years at daily resolution (for Europe see e.g. Mediero et al. 2015), which means that the extreme events of interest in this study, i.e. those with return periods of > 50 years are very rare in observed records—and a sample size large enough to robustly estimate*

the magnitude of rare events with return intervals of 50-200 years simply does not exist. These inherent limitations of the historical observational record can be obviated through the use of a climate model large ensemble approach (van der Wiel et al. 2019) in combination with an advanced hydrological model as used in our study. We highlight the limitations of observational records and the need to use a hydro SMILE to study very extreme events in Section 1.1. (from l. 68): 'Addressing the precipitation-flood paradox is simply not possible using observations alone, as the high-end extreme events of interest are rare to non-existent in temporally limited observational records. This real-world data limitation effectively precludes statistical analyses of extreme events with return periods exceeding ~50 years. To overcome this problem, we use a hydro-SMILE to obtain a large number of extreme P-Q pairs.'

(4) It seems that the term 'extremeness' is another term for return interval (line 68). I think that it would be more intuitive to simply use the return interval rather than invent a new (and somewhat awkward) word. It would be simple to say that there exists a "return interval threshold" or a "threshold based on the return interval."
Reply: *Thank you very much for pointing out the need for rephrasing. We replaced the term 'extremeness threshold' by 'return interval threshold' everywhere in the manuscript except in two instances in the abstract and the introduction where we associate the term 'return interval' with 'extremeness' for those people who might be less familiar with return intervals.*
Modification: *several instances in the manuscript*

References used in the responses to the reviewers

Addor, N., Rössler, O., Köplin, N., Huss, M., Weingartner, R., & Seibert, J. (2014). Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments. *Water Resources Research*, *50*, 1–22. https://doi.org/10.1002/2014WR015549

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. *Climate Dynamics*, *38*(3–4), 527–546. https://doi.org/10.1007/s00382-010-0977-x

Hattermann, F. F., Vetter, T., Breuer, L., Su, B., Daggupati, P., Donnelly, C., Fekete, B., Florke, F., Gosling, S. N., Hoffmann, P., Liersch, S., Masaki, Y., Motovilov, Y., Muller, C., Samaniego, L., Stacke, T., Wada, Y., Yang, T., & Krysnaova, V. (2018). Sources of uncertainty in hydrological climate impact assessment: A cross-scale study. *Environmental Research Letters*, *13*(1). https://doi.org/10.1088/1748-9326/aa9938

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., & Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, *11*(2), 491–508. https://doi.org/10.5194/esd-11-491-2020

Mediero, L., Kjeldsen, T. R., Macdonald, N., Kohnova, S., Merz, B., Vorogushyn, S., Wilson, D., Alburquerque, T., Blöschl, G., Bogdanowicz, E., Castellarin, A., Hall, J., Kobold, M., Kriauciuniene, J., Lang, M., Madsen, H., Onuşluel Gül, G., Perdigão, R. A. P., Roald, L. A., … Orarinsson, O. (2015). Identification of coherent flood regions across Europe by using the longest streamflow records. *Journal of Hydrology*, *528*, 341–360. https://doi.org/10.1016/j.jhydrol.2015.06.016

Prein, A. F., & Pendergrass, A. G. (2019). Can we constrain uncertainty in hydrologic cycle projections? *Geophysical Research Letters*, *46*(7), 3911–3916. https://doi.org/10.1029/2018GL081529 van der Wiel, K., Wanders, N., Selten, F. M., & Bierkens, M. F. P. (2019). Added value of large ensemble simulations for assessing extreme river discharge in a 2 °C warmer world. *Geophysical Research Letters*, *46*(4), 2093–2102. https://doi.org/10.1029/2019GL081967

Vetter, T., Reinhardt, J., Flörke, M., van Griensven, A., Hattermann, F., Huang, S., Koch, H., Pechlivanidis, I. G., Plötner, S., Seidou, O., Su, B., Vervoort, R. W., & Krysanova, V. (2017). Evaluation of sources of uncertainty in projected hydrological changes under climate change in 12 large-scale river basins. *Climatic Change*, *141*(3), 419–433. https://doi.org/10.1007/s10584-016-1794-y

7th Jul 21

Dear Dr Brunner,

Please allow me to apologies for the delay in sending a decision on your revised manuscript titled "Reconciling the extreme precipitation-flood paradox in a warming climate". It has now been seen by our original reviewers; reviewer 1 just recommended publication and had no further comments, reviewer 2's 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 edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

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 https://www.nature.com/commsenv/about/open-access

At acceptance, the corresponding author will be required to complete an Open Access Licence to Publish on behalf of all authors, declare that all required third party permissions have been obtained and provide billing information in order to pay the article-processing charge (APC) via credit card or invoice.

Please note that your paper cannot be sent for typesetting to our production team until we have received these pieces of information; therefore, please ensure that you have this information ready when submitting the final version of your manuscript.

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,

Heike Langenberg, PhD

Chief Editor
Communications Earth and Environment

On Twitter: @CommsEarth

REVIEWERS' COMMENTS:

Reviewer #2 (Remarks to the Author):

I have now read the reviewer comments and the revised manuscript. I appreciate the authors' thoughtful responses to my comments and I appreciated their explanations and revisions, which helped me to clarify my understanding of the work and deepened my appreciation for the breadth of the analysis and modeling that was completed. At this point, I am in agreement that the major editorial concerns have been addressed and I look forward to seeing the manuscript published. I do also look forward to the authors expanding their modeling work to other areas, as I believe their approach has potential to shed light on this long-standing conundrum between changes to precipitation and flooding.

Reviewer #2
I have now read the reviewer comments and the revised manuscript. I appreciate the authors' thoughtful responses to my comments and I appreciated their explanations and revisions, which helped me to clarify my understanding of the work and deepened my appreciation for the breadth of the analysis and modeling that was completed. At this point, I am in agreement that the major editorial concerns have been addressed and I look forward to seeing the manuscript published. I do also look forward to the authors expanding their modeling work to other areas, as I believe their approach has potential to shed light on this long-standing conundrum between changes to precipitation and flooding.

Reply: *We again thank the reviewer for their constructive inputs, which helped to clarify certain methodological aspects of our work and to strengthen the storyline.*