Response to reviewer comments on "Widespread flooding dynamics changing under climate change: characterising floods using UKCP18"

The authors would like to thank the 3 reviewers and the Editor for their further comments, and their perseverance in helping us improve the manuscript. We describe below how we have addressed each comment, and sincerely hope that the reviewers and Editor now consider that we have adequately dealt with all of their comments and that the manuscript is now suitable for publication.

Editor:

From my reading of the manuscript as an Editor I concur with reviewers #1 and #3. I encourage the authors to revisit their approach along the lines of the suggestions of reviewers #1 and #3.
Please see our response to each reviewer comment below.

I find it disappointing that the authors did not address my own comments. I still think that an article published in an international journal such as HESS must be of interest to readers outside the authors' country, e.g. by reporting methodological advances. I am sorry to say that the authors did nothing to emphasise the method over the UK-centric implications. They conclude that the number of widespread events in the UK and their seasonality will increase. Why would this be of interest to anyone outside the UK? Without an attempt to address this concern, the paper will unlikely become acceptable.
We have now made a number of changes, with the aim of making the manuscript of greater interest internationally and less UK-centric. These include

- Changing the title to make it more general.
- Adding more about the method to the abstract, as well as emphasising that the methods/analyses could be applied elsewhere.
- Edits to the Introduction to make it more general, and again emphasise that the methods/analyses could be applied elsewhere.
- Separating the Discussion into a part relating to the GB results and a part relating to the methodology, with the latter now including discussion of what particular aspects of the method may need to be specifically considered for applications in other regions.
- The Conclusions now stating "While the focus here was GB, the methods and analyses described could be applied to other regions with hydrological models and climate projections of appropriate resolution. The most suitable definition for a 'widespread flood event' may vary for other countries/regions."

We hope that these edits now make the manuscript of much greater interest internationally.

Reviewer 1:

Overall comments and recommendations

The authors have given this revision of the manuscript considerable thought and effort, and the additional material helps to highlight both the challenges and value of the work undertaken. I raised two major concerns previously, namely 1) the need to assess the efficacy of the model simulations against suitable reference data, and 2) the manner in which the probabilities of exceedance were estimated. I am comfortable with the changes made to address the latter concern, so my comments below are focused on the efficacy of the adopted modelling chain. I sat on my draft comments for a week before finalising as I would like to see the paper published and I appreciate that the authors will be getting frustrated with this review process; however, I am struggling to agree with the author's interpretation of the new evidence presented and to date only the RCM projections and not the G2G model itself has been assessed.
Thank you for your continuing efforts to help improve our manuscript. We hope that the substantial edits and additions now made, described below, are sufficient to justify acceptance/publication.

The key contribution of the paper rests on characterising the possible changes to areal flood behaviour under climate change, where the adopted method implicitly accounts for the joint interactions between soil moisture, rainfall, and the non-linear influence of increasing catchment scale. The work leverages a large body of prior work undertaken to provide the adopted regional climate model (RCM) ensembles and G2G (hydrologic) modelling; while the use of this inherited work is of great benefit, it also brings with some specific challenges relevant to the stated research objective.

It is true that the work reported in this manuscript relies upon a large amount of preceding work, which has both advantages and disadvantages. It also meant that, in the interests of brevity, we were trying not to repeat too much from elsewhere – a difficult balance, that we got a bit wrong but have hopefully now improved!

My interpretation of the new information provided in the last revision is that the RCM ensembles do a poor job of simulating the temporal and spatial correlations of the key processes of interest, and this undermines the core conclusions of the paper as currently framed. It also raises questions about the defensibility of the G2G simulations. While it is perhaps unreasonable to expect that the authors somehow "fix" deficiencies in the prior modelling effort that they have inherited, I do think it important that any shortcomings in the modelling chain are explored and used to qualify the conclusions made and to highlight areas for future research.

The differences in flood characteristics between the baseline climate model-driven run and observation-driven run have now been described more fully in the Results (Section 4), and a paragraph has been added to the Discussion (end of Section 5.1) discussing the differences and the possible reasons for the them, and areas for future work. The need to interpret future changes "in the context of any differences in event characteristics between the baseline climate projection-driven model runs and an observation-driven model run" has also been added to the Abstract and Conclusions (Section 6 para' 2).

It appears that my assessment of the new information differs from the stated views of the authors, and this may be because I am misinterpreting information presented in the manuscript. So, either my concerns point to the need for the authors to better clarify and justify their reasoning, or else to perhaps revise their views based on the comments below.

I think the contribution of the paper would be greatly enhanced if the authors:

1) replace the current Figure 1b with an equivalent figure based on G2G results forced by "observation-based simulations" (rather than by the current 12-member RCM ensemble), and replace Figure 1a with a simple scatter plot of observed vs simulated Q50; the former plot provides insights about spatial biases related to regional hydroclimatic differences, and the latter plot about the nature of any overall model biases.
2) adopt a more critical approach to reviewing the implications of the results from the proposed G2G assessment as described in the previous point, and also from the RCM ensemble assessment already undertaken and currently summarised in Figures 4, 6 and 7.

The effort involved in undertaking 1) should be modest as the necessary hydrologic simulations have already been undertaken, but if this is not possible for some reason, then the manner in which the discussion and conclusions are written could be revised to better highlight, or refute, the perceived limitations in the hydroclimatic projections on which the main conclusions are based.

1) The suggested replacement for Figure 1b (a map showing the flood performance of the SIMOBS run) is available elsewhere (Kay 2022) - we have now added a paragraph describing these results and some older work that shows similar patterns of performance (Section 3.1 para' 3) (unfortunately an oversight meant that the older work was not cited in this manuscript before). The recent work also includes maps of performance with and without a simple bias correction of

precipitation – information on this has been added (Section 2.2 para' 2), and Fig 1a replaced with the suggested scatter plot (Section 3.1).

2) As described above, a paragraph has been added to the Discussion about the differences between the SIMOBS and baseline RCM-driven results, the possible reasons for the them, and areas for future work. The need to interpret future changes "in the context of any differences in event characteristics between the baseline climate projection-driven model runs and an observation-driven model run" has also been added to the Abstract and Conclusions (Section 6 para' 2).

Further rationale for these recommendations is provided below.

RCM projections

Comparing the 1981-2010 flows derived from the RCM ensemble with those from the "observation-driven" inputs provides a very effective evaluation of the modelled climate over the 1981-2010 period. It is more valuable than comparing any individual climate variable alone (e.g. rainfall) as it implicitly allows for the joint distribution of climatic factors that influence floods, so this is a very useful addition to the paper.

However, at present the authors conclude that the RCM simulations of event areas are "fairly consistent" with those derived using SIMOBS, and that there is "slight bias" in the distributions of event severities. I find it difficult to accept these conclusions as stated because:

1) the differences in results between the SIMOBS and the mean RCM ensemble simulations as shown in Fig 4 are appreciably larger than the modelled differences between the climatic baseline and future conditions, and

2) the SIMOBS results appear to lie outside the maximum and minimum range of individual ensemble RCM results over the period 1981-2010 as deduced from a comparison between Fig 4 of the MS and Figure 1 of Supplementary material.

The differences in results would be better illustrated by including 'error bars' in Fig 4 that show the min/max spread of ensemble results for the 1980-2010 period, or else illustrating these differences in some other more easily comprehended fashion. It would also appear that the RCM simulations greatly overestimate the duration of the events (Fig 6), where it is observed that the differences between the RCM ensembles and the SIMOBS results are considerably greater than the projected differences due to global warming. The simulation of flood extents (Fig 7) appear more reasonable (but there is still an order of magnitude difference between simulated and observed maxima), and I agree with the interpretations of the authors regarding the adequacy of the seasonality of the flood regime as shown in Fig 5.

Error bars have been added to Fig 4, in a similar way to Fig 5.

It is difficult to assess how well the heatmaps for SIMOBS vs baseline RCM-driven compare (Figs 6, 7), due to the much more limited number of events available from the SIMOBS run than the pooled baseline RCM-driven runs. This is particularly the case when considering combinations of properties (as for the heatmaps, rather than the preceding histograms), and particularly for more extreme combinations (as is pointed out in the Discussion, Section 5.1 para' 1).

As described above, the differences in flood characteristics between the baseline climate model-driven run and SIMOBS have now been described more fully in the Results (Section 4) and a paragraph has been added to the Discussion (end of Section 5.1).

In short, the results shown in Figures 4, 6 and 7 appear to suggest that the RCM climate ensembles overestimate the spatial dependence of rainfall events and over-estimate the serial dependence in the correlation structure of rainfall extremes, and further that the distribution of modelled areal maxima is perhaps not as "fat tailed" as the observed data. Given these characteristics are of core importance to the main conclusions of the paper, I think these issues need further discussion.

As described above, a paragraph has been added to the Discussion (end of Section 5.1) about the differences between the SIMOBS and baseline RCM-driven results, the possible reasons for the them, and areas for future work.

G2G modelling

In their most recent response to reviewers, the authors state that "a direct comparison with observations is not possible" as gridded observations of flow do not exist. This is an oddly naïve statement to make as first, this is obviously the case everywhere in the world, and second, gridded model outputs can be aggregated to represent catchment runoff and compared to observations at selected streamflow gauges; also, if I interpret Figure 1 correctly, isn't this exactly what the authors have done to investigate model bias in estimates of daily 50-year return period (Q50) flows? That is, isn't Figure 1 based on comparing aggregated gridded runoff outputs (or routed flows) with locations where streamflow gauges are available?

The outputs from the G2G model are in fact river flow (routed runoff), not grid-cell runoff – this has been clarified (Section 3.1 para' 1). Indeed, previous work has evaluated flow simulations from G2G against gauged flows for locations across GB, and some more recent work on this has now been described/cited (Section 3.1 para' 3). The point we were trying to make when stating "a direct comparison with observations is not possible" was relating to features like flood extent and coherence, which cannot be properly investigated using gauged flow data for a greatly more limited set of river locations than are available from our gridded model outputs.

My concerns with the current Figure 1 are that no mention is made of what bias-correction method was used (simple delta scaling, quantile scaling, or something more sophisticated? was it applied to the climate projections or directly to the Q50 estimates?); but more importantly, it lumps the potential biases of both the RCM climate projections and the G2G modelling together, so it is not possible to determine which aspect of the modelling chain is causing problems, and thus how the biases in either (or both) the G2G and RCM modelling should best be addressed. Accordingly, the more useful approach would be to compare the Q50 results derived from SIMOBS results to the Q50 estimates derived from the gauged streamflows as this then reveals the efficacy of the G2G model separately from the efficacy of the RCM simulations, which are already explored in Figs 4 to 7.

Reference has been added to previous work showing SIMOBS flood peak performance vs gauged data (Section 3.1 para' 3). A discussion of bias correction has been added (Section 2.2 para' 2), and Fig 1 has been replaced with a scatter plot (as suggested above).

I assume that the authors' justification for the adequacy of the G2G modelling relies on the basis of previous published papers. I have thus gone back and searched the literature cited in the paper (noting that the Kay et al. 2018 was missing from the reference list) but I could not find any assessment in the published papers of the G2G model's ability to estimate the extreme floods of most relevance to this paper. On the basis of the model's formulation, I would expect that the G2G model is well suited (if appropriately parameterised) to characterising low and high flow regimes and soil moisture behaviour, but that it would struggle to represent the extreme floods of interest. Without seeing evidence to the contrary, I would expect that the G2G model would be better suited to water resource applications rather than flood applications as the model structure and parameterisation is focussed on the gross partitioning of rainfall, evapotranspiration, soil moisture accounting, and the redistribution of sub-surface moisture; while these state variables are relevant to antecedent conditions which influence flood behaviour, they are not well suited to characterising the flood response during an extreme event. From a purely information content perspective, 99.98% to 99.99% of the daily data used to inform the G2G parameters relate to non-extreme conditions. Unless special steps are taken to use the ~0.02% of information relevant to extreme flood behaviour (and the model structure and parameterisation is able to take advantage of it), then it is unlikely that such a model is able to adequately represent flood conditions. I would expect such models tend to

provide precise information at highly resolved spatial and temporal scales, but the estimates are likely to be biased and inaccurate, particularly when the model has been configured to provide estimates at regional/national scales rather than catchment. I do understand the value and trade-offs involved in developing national-scale models as opposed to catchment-specific models, but I also think we need to be transparent about how the performance of such modelling schemes.
Apologies – this was an oversight in the original manuscript, which omitted references specifically looking at flood estimation. A paragraph has been added describing this and other more recent work (Section 3.1 para' 3). In addition, further information and references have been added regarding the operational use of G2G for flood forecasting across England, Wales and Scotland (Section 3.1 para' 2).

Of course I may not be right in this instance as my expectations are based on my own experience of using a range of conceptual and "physically based" models, and not of this G2G model. However, my difficulty here is that I could not find any evidence in the current manuscript, or in previously published papers, that the G2G model is able to provide reasonable estimates of the Q50 flood. My above recommendation for comparing the Q50 estimates derived using the SIMOBS inputs would provide clear evidence as to how well the model is performing at the national scale of interest, and these insights can be used to add defensibility, or caveats, to the conclusions drawn.
Thank you for pointing out the omission, which we hope we have now rectified.

Reviewer 2:
Given the previous round of feedback provided by myself and the other reviewers, I had expected a much more substantial revision in terms of clarifying data and methods, validation of G2G, and subsequent discussions. I'm afraid that this revision suggests that the authors have not grasped the magnitude of verifications needed to ground their study, how these would inform the interpretation of the results, or justify the conclusions they wish to make. I still find that the motivation of this study is well-founded but it has not been adequately executed (and the implications of their results have not been clearly communicated in the abstract or conclusions – who would benefit from this information? Does it have the potential to alter current practices?) and it may be better for the research process for me to reject the manuscript and allow the authors space to reassess their approach and perceptions of the results without being tied to the manuscript in its current form. The authors have partially addressed the need, and made a significant effort, to provide evidence that the flood characteristics modelled using baseline climate is representative of floods modelled using observed historical climate and I commend them for undertaking this effort. However, this analysis step, which I see as critical to establishing the credibility of the subsequent analyses, is somewhat concealed and is, at present, still inadequate.
We have now made a number of substantial additions, to

1. clarify data/methods/G2G evaluation (Sections 2 and 3),
2. more fully discuss the results and their implications (Sections 4 and 5), and
3. broaden the Abstract, Introduction and Conclusions (Sections 1 and 6).

We hope that these revisions are now sufficient to meet your concerns.

The data inputs used to drive the SIMOBS runs are not described under section 2 of the data, but rather briefly introduced in Section 3 of the methods. I had previously assumed that G2G was run at a daily time step given the RCM outputs were at a daily time step. (Note that details regarding the time step of the flood modelling are missing and need to be rectified. Apologies for missing this before.) However, the description of the observed data states that "Precipitation was subdivided uniformly through the day, and temperature varied sinusoidally between the extremes." suggesting that G2G was modelled at a subdaily time step for the SIMOBS runs. This is problematic because the floods modelled from SIMOBS and the baseline runs will not be comparable. Assuming a uniform

subdaily rainfall pattern (if G2G is intended for producing subdaily flood responses) introduces other errors as a uniform rainfall pattern is much more likely to result in a smaller flood response compared with a more variable rainfall temporal pattern. The G2G runs based on data that's been gridded based on observations needs to equivalent in time steps and spatial resolution to the baseline runs.

A sub-section has been added (new Section 2.1), describing the observation-based driving data. Information on the model time-step has been added (Section 3.1 para' 1). Both the SIMOBS and SIMRCM runs are done in the same way, e.g. by equally sub-dividing daily precipitation data – this has been clarified (Section 2.2). Recent work (Kay and Brown 2023) has shown that, while use of hourly, rather than equally-distributed daily, precipitation does improve the simulation of flood peaks derived from daily mean flows, it has relatively little effect on the simulated future changes in peak flows for the larger catchments (>= 50km2) included in this work. Information on this has been added (Section 3.1 para' 3). We typically do not include small catchments, because of the use of daily data (now clarified in Section 3.1 para' 1). The potential for differences in the SIMOBS and SIMRCM results to be due to differences in spatial resolution of the respective driving data has been added to the Discussion (Section 5.1 para' 4).

The authors provide references that provide evidence that the G2G hydrological model has been verified for the purpose of analysing floods across Great Britain and for flood forecasting. If these studies specifically reproduce the specific flood features of interest in this study, namely spatial extent and coherence, seasonality, frequency, and duration, then this needs to be stated and will be the validation needed for this study. If the references provided by the authors do not specifically validate the reproduction of the flood features relevant to this study, then this validation needs to be conducted by the authors here for any of the subsequent analysis to be credible. I acknowledge the challenge in doing this given the authors have pointed out the difficulty in validating G2G outputs based on observed climate data, as they have now modelled, given the lack of gridded flow data based on observed streamflow data. However, validations of gridded runoff and streamflow products are not at all without precedence, for example: Frost, A.J., Shokri, A., Keir, G., Bahramian, K., Azarnivand, A., 2020. Evaluation of the Australian Landscape Water Balance model (AWRA-L v7) (Bureau of Meteorology Technical Report) and I note that Reviewer 2 had suggested that such an evaluation could be conducted at some representative locations and this too would be sufficient.

A number of references relating to previous work evaluating G2G for flood modelling and forecasting have now been added, with further detail (Section 3.1 para's 2,3) – apologies that most were initially omitted. However, they do generally focus on evaluation against gauged flow data at a set of individual gauge locations. We do not believe that it is possible to specifically evaluate factors like spatial extent and coherence of floods in this way, given the very limited number of gauged locations (<800) relative to the number of 1km river grid cells across GB (nearly 20,000). Further barriers/complications to such an evaluation include differing periods of data availability, differing reliability of high flow gauging, and differing artificial influences on gauged flows.

Having looked at the BoM report cited above, we cannot see where this looks at evaluating spatial extent or coherence of floods (but apologies if this has been missed). It may be possible to use remotely sensed spatial datasets of, for example, soil moisture to help evaluate spatial performance of a model, but even then there are difficulties in terms of being able to compare equivalent things, as remotely sensed data typically only include moisture in the top layer of the soil, whereas G2G provides depth-integrated soil moisture for the whole soil column.

In terms of spatial extent and coherence of floods, it is the use of the underpinning spatial datasets on landscape properties that achieves this in combination with the spatial pattern of precipitation over time. The grid-based approach does not aggregate landscape properties at the scale of the gauged catchment, nor does it employ catchment-average rainfall, so the spatial extent and coherence of flooding should be better captured across the modelled domain. Figure 2 of Moore et al. (2012) provides an insightful illustration of the extent and coherence properties of G2G flood risk

assessments for a 10-year flood exceedance for an area over the English Midlands during the summer 2007 floods. Also, Moore et al. (2006) specifically considers the issue of forecasting extreme floods and the value of a distributed model such as G2G in this context, including illustrative case studies. A detailed evaluation of G2G for use in rapid response catchments (typically ungauged) is made in Cole et al. (2013). This provides evidence, across Britain over four water years and for case study storms, of good performance stratified by catchment type (area, urbanisation, headwater).

The results and discussion need to be presented with respect to the validation results. The ability of the SIMOBS to reproduce the flood features of interest will determine the caveats with which the climate change impact results are interpreted. At present, given the validation of SIMOBS in reproducing these flood features are missing, I find that the discussion and conclusions do not adequately represent the level of conjecture that should currently be ascribed.
As stated in response to similar comments from Reviewer 1, the differences in flood characteristics between the baseline climate model-driven run and observation-driven run have now been described more fully in the Results (Section 4), and a paragraph has been added to the Discussion (Section 5.1 para' 4). The need to interpret future changes "in the context of differences in event characteristics between the baseline climate projection-driven model runs and an observation-driven model run" has also now been added to the Abstract and Conclusions (Section 6 para' 2). We hope that these changes make the level of uncertainty much clearer.

I strongly encourage the authors to revisit their approach as their aim to examine the impacts of climate change on widespread flooding is a worthy endeavour. I can see that they have undertaken substantial analyses to shed light on this topic and these efforts should not be cast aside. Rather additional work is needed to properly substantiate their claims. This work contains the endings of what could be a very good paper, but the foundations need to be properly established.
Thank you for your continuing efforts to help improve our manuscript. We hope that the substantial additions now made, described above, are sufficient to justify acceptance/publication.

Reviewer 3:
I want to thank the authors for taking the time to carefully address the comments made during the first round of revision. It has greatly improved in my opinion.
Thank you.

Some final suggestions for technical corrections:
- Add an explanation of the percentages shown in the figures to the caption of Figure 3.
Done (moved from main text).

- Figure 4 is difficult to read. I suggest to use a histogram style similar to the one presented in Figure 5.
Done.

- x-labels are missing in Figure 9
Done.