Response to Reviewers and Editor

Responses to the reviewers are given in blue following each point. Line numbers correspond to the first revision, not the new version.

Reviewer 1

General Comments

The motivation of the paper to examine the spatial and temporal coherence of flood risks is well justified and the authors' approach to examining this topic is efficient and effective. However, the justification for using raw climate data for examining floods resulting from extreme climates is not yet satisfactorily justified, but I believe an additional validation step would allow the authors to retain the approach they have already adopted. I find the manuscript to be overall well written and figures are mostly well presented (some improvements are needed for Figs 3, 4, 7, and 8 and Figure 5 needs to be revised). A number of corrections and clarifications are necessary to ensure that: consistent terminology is used; the language considers the global readership of this journal; descriptions are precise and objective; figures are easily interpreted; and referencing is accurate.

Major comment

I concur with the authors that the use of bias correcting in an attempt to capture rainfall extremes relevant to floods is difficult and introduces considerable uncertainty and the justification here is to avoid the introduced uncertainty and instead examine the differences in floods resulting from modelled historical and future climates. However, the justifications currently provided are inadequate and at times illogical. The authors state that "due to the focus on [sic] the present work on extremes rather than the whole regime in general, bias correction is not applied here."

I'm afraid this justification is unsatisfactory. I'm certain the authors are knowledgeable of the fact that regional climate models are aimed at resolving large scale spatial and temporal variability, namely, what is referred to here as "the whole regime in general". A study that was aimed at large scale changes would therefore be well justified in using the raw RCM outputs. In contrast, extreme precipitation events are typically driven by synoptic scale events, which would justify the use of bias-corrected projections. Furthermore, flooding is sensitive to the spatial and temporal distribution of storms at even finer scales and have the added complexity of local factors that influence the flood response other than climate. The focus on extreme events is therefore a justification for introducing uncertainty to address the increased complexity of flood responses. Lines 65-66 therefore needs to be removed.

Lines 65-66 have been removed. Instead the following has been added at the end of Section 3.1

An investigation was undertaken to identify whether bias correction should be used in this paper. G2G outputs based on the UKCP18 RCM ensemble members was compared to daily mean flow (data available from the NRFA). The 50-year event (annual exceedance probability of 2%) was calculated for the station and the relevant gridsquare it lies in. Figure 1 shows that across most of Great Britain where the mode was run, bias correction led to a fairly constant underestimation of the 50-year event compared to those from observations. Although the results without bias correction are more variable with some stations showing a large overestimate, they have a better mean bias when calculated nationally, which was felt to be important when looking on a national scale. This was also computed for the 2-year flood with very similar results. Due to this, it was decided that bias correction would not be applied in this paper.

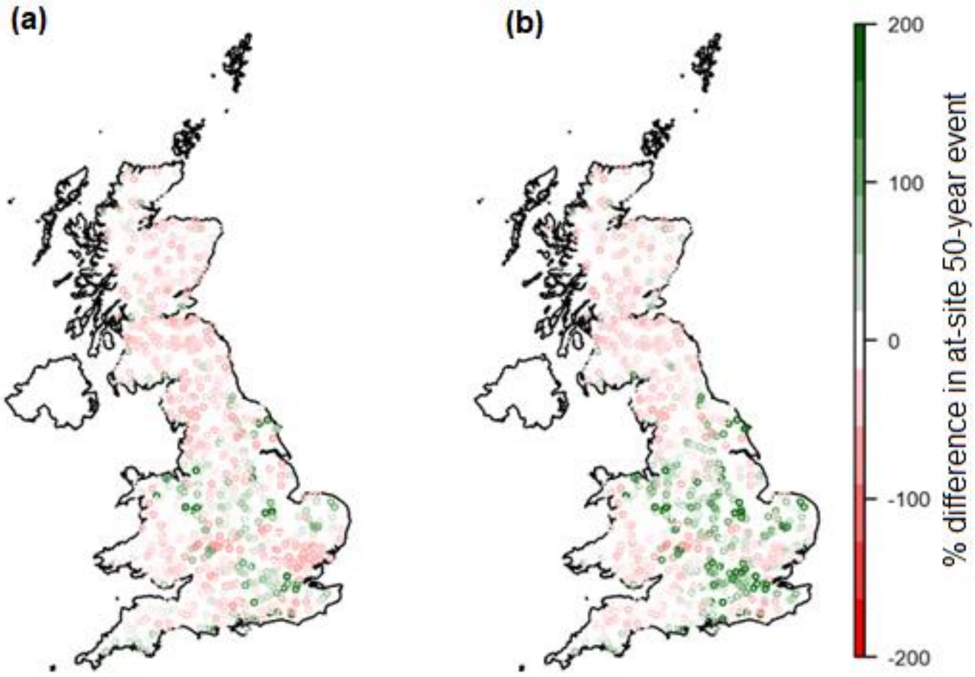

FIGURE 1 COMPARISION OF DIFFERENCE BETWEEN (A) WITH AND (B) WITHOUT BIAS CORRECTION AVERAGED OVER ALL 12 ENSEMBLE MEMBERS. COLOUR INDICATED THE CHANGE BETWEEN THE ESTIMATE FOR THE 50-YEAR RETURN PERIOD PEAK FLOW (BASED ON GAUGED DAILY FLOW). POSITIVE VALUES INDICATE THAT MODELLED DATA HAS A LARGER VALUE OF Q50.

In their response letter, the authors make the case that they have kept the biases constant by comparing modelled historical and modelled projected floods and by capitalising on the dynamic nature of the GCMs. (however, it then follows that the same justification would not rule out the comparison of bias-corrected data – here too, the biases would be consistent between modelled historical and projected climates. Furthermore, bias corrected rainfalls are not constrained to the observation based datasets, as it appears to be suggested in the authors' response– both statistical and dynamical downscaling approaches allow for estimates outside the range of the observed records.)

This justification is sufficient and acceptable conditional upon the authors adopting the following suggestions, which should be easy to implement:

1. Repeat the analysis using observational data over the same period as the historical modelled data to quantify the degree to which the spatial and temporal coherence of flooding is approximated and represented in the baseline case. This would then provide evidence of the fidelity of the baseline to which the relative changes in the number of widespread events could be estimated using the approach that has already been presented. Currently, the presentation of results dependent only on modelled climate data makes it easy for a reader to suppose that the results reflect similarities in catchment characteristics combined with the lack of spatial specificity in the modelled climate date. Having an observation-based analysis would provide context. Presenting these observed results in the SI would be adequate

2. Present the findings as relative changes in the number and spatial scale of the widespread events (more details of this in reference to figure 3 below)

We thank the reviewers in their continued interest in this point. Unfortunately, gridded observations of flow do not exist nationally for the United Kingdom, and so a direct comparison with with observations is not possible. However, we do now include a comparison with some "observation-driven" simulations (SIMOBS), which have been discussed and analysed previously in Kay (2022). These observation-driven simulations make use of observed gridded rainfall, gridded potential evapotranspiration and gridded daily observed temperatures. We add panels or bars to Figures 3, 4, 5, and 6 which show the values for the SIMOBS run (which only exists for the baseline timeslice).

The following is added to the methods section:

In isolation it is difficult to say whether these results are realistic compared to observations. However, gridded river flow observations are not available for Great Britain nationally. Therefore, in addition to the UKCP18-driven G2G output, this paper also presents data from a set of “observation-based simulations” as used in Kay (2022) as a step towards comparing modelled and observed extreme flow. This run still uses Grid-to-Grid but is driven using observed inputs: CEH-GEAR daily gridded precipitation (Tanguy et al., 2016), monthly short grass potential evapotranspiration (40km resolution) from MORECS (Hough and Jones, 1997), and daily 1km minimum and maximum daily temperatures (Met Office et al., 2019). Precipitation was subdivided uniformly through the day, and temperature varied sinusoidally between the extremes. In this paper, this will be referred to as the SIMOBS run.

The following is added at line 194 in reference to Fig 3

Fig 3 shows that the event areas are fairly consistent between the RCM-driven runs and the SIMOBS run, with a slight bias in the RCM-driven runs to larger events with lower return periods. All the RCM-driven runs show a slightly flatter distribution of return periods in the 2050-2080 time-slice.

The previous version of Figure 3 (with some minor formatting changes) has been moved to Supplementary Material Figure 1.

The following is added at line 211 in reference to Fig 4

In comparison, the SIMOBS run shows a more equal distribution of events across the seasons, though it still remains within the variability of the seasonal totals for the RCM-based baseline outputs except in autumn (SON).

The following is added to line 225 in reference to Fig 5.

The SIMOBS run appears to generate shorter events on average compared to the RCM-driven runs, suggesting a slightly stronger temporal autocorrelation in the effects of the use of UKCP18 input data. The return periods (as seen in Figure 3) are broadly similar in distribution.

The following is added to line 232 in reference to Fig 6.

The SIMOBS run shows a broadly similar distribution to the baseline (1980-2010) timeslice, although the variability and reduced smoothness appears to be increased, due to the much smaller number of events from that single run (~500 compared to ~7000 from all 12 RCM ensemble members).

Specific comments (line numbers reference manuscript 2)

Acronyms are often used before they are defined – the ones I noticed were: NRFA, RCM (this shouldn’t be assumed knowledge since GCM is previously defined), PoE. There may be others.

All abbreviations spelled out on first usage.

It would be helpful to be clarify the terminology used to describe seasons - this is primarily for your readers in the southern hemisphere, please don’t neglect us. Please consider including “boreal” when seasons are referenced. Alternatively, include the months in brackets after each season. It may seem superfluous for a northern hemisphere native, but it makes interpreting the results so much easier for those in the south.

Months are referenced when seasons are mentioned, e.g. spring (March-May), summer (June-Aug), autumn (Sep-Nov) and winter (Dec-Feb).

The time slices are introduced as a “baseline” and “future”, but the former is later referred to as “present”. Please keep the terminology consistent, particularly since “present” is inaccurate when referring to a time period covering 1980-2010.

We agree with this comment. All references to “present” are swapped with “baseline”.

L 90: The term “percentiles” in reference to floods is used to describe distributions. The term that should be used here is frequencies, not percentiles.

We change this to “frequencies”.

L95: similarly, the notation Qx is widely used in hydrology to denote the x percentile of flow rather than a flood frequency. Please keep the terminology you have chosen consistent by referring to the 1 in 5 and 1 in 10 year probability event as POT 0.2 and POT 0.1 respectively.

Alternatively, use the term Average Recurrence Interval of 5/10 years (ARI5 and ARI10).

We will use POT0.2 and POT0.1 throughout instead of Q5 and Q10.

L99: This is ambiguous as absolute values of flow are in fact being used as thresholds, and these thresholds are dependent on the distribution of the data: the definition of the thresholds are just location dependent. I suggest phrasing this as:

Note that the thresholds are not based on universally applied fixed values of flow magnitude but are instead dependent on thresholds defined by empirical flood frequencies.

Thank you for this suggestion. We replace the sentence with the one offered.

L150: I had to read this several times and I’m still unclear. The text implies that $\bar{\chi}$ applies to large flows and conversely that χ applies to flows of all magnitudes (which I do not believe is the intended message). I suggest the following (but I’m unsure whether I’ve interpreted the text correctly. Please revise as necessary)

χ describes the level of asymptotic dependence; if χ^{1}0 then the variables are asymptotically dependent. A value of χ=0 represents asymptotic independence. Asymptotically independence is also represented by a value of $\bar{\chi}^{1}$1. A value of $\bar{\chi}$=1 represents flows that are dependent but nor asymptotic.

We replace the existing paragraph at line 150 with: χ describes the level of asymptotic dependence; if $\chi > 0$ then the variables are asymptotically dependent, and $\bar{\chi} = 1$ automatically. But if $\chi = 0$, they are asymptotically independent. In this case, $\bar{\chi}$ describes the dependence for large but not asymptotic values of flow. $\bar{\chi}$ close to 1 indicates the variables are highly dependent except at the asymptotic limit.

L185: please justify why these four events are selected. Why “four of” and why not “the four largest”? Are they selected to demonstrate that most events are spatially contiguous? (As an aside to the major comment above, is the spatial coherence an attribute derived from the modelled climate data, or is the pattern present in observed data?)

We correct line 176 to say “Fig 2 shows the four events with the widest spatial extent…” Due to a lack of gridded observed flow data available for the United Kingdom, it is not possible to compare these events with observed events.

Figure 2: If I’m understanding this correctly, this figure shows the degree of spatial inundation with the colour scale showing the equivalent severity of the flood for events identified using the POT2 threshold, equivalent to a return period of 0.5. If this is the case, why are more frequent events with return periods of 0.2 to 0.5 shown? Perhaps they are not, but the light yellow scale is difficult to distinguish between greater or less than 0.5. I suspect it is simply a case of the legend needing to be updated to show grey between 0.2 and 0.5. Also, the text size of the labels on this and the next figure are rather disproportionately large. Please reduce the text size.

The reviewers are correct – the light yellow should be coloured grey, no points on the map are actually coloured using that shade. We also correct the label sizes.

Figure 3: Comparing the changes between the baseline and future is not easy with the way this figure is presented – there are a lot of vertical bars in different ensemble members to compare and lining up the number of events for different return periods is challenging, while the colour selection implies that A and B are showing a different variable to C and D. I strongly suggest combining the information from A and C, and B and D by showing these results as the difference between time slices. If there is a very good reason to not do this, please at the very least change the colours to paired colours: e.g. light blue for A, blue for C, light green for B

and green for D, so that light colours correspond to the baseline, darker colours to the future, blues for area, and greens for return period.

To compare with an additional run using "observation-driven simulations (SIMOBS)" as discussed above, Figure 3 has been moved to the supplementary material, and replaced with a figure that compares the SIMOBS with a weighted average of the ensemble members for the baseline timeslice and the future timeslice (weighted by number of events extracted), which scales the number of events to be comparable to the SIMOBS run. This combines panels A/C and panels B/D, using shading for the ensemble average and two outlines for present and future.

L 203-2007: The choice of using the word "may" makes the sentence sound speculative. These sentences could be revised to reflect the certainty of the results. Suggest the following or similar:

However, the increase in widespread events is confined to the boreal autumn (SON) and winter (DJF) with a decrease in events between March and August. The decrease in future boreal spring and summer events could be due to overall projections of drier summers (Murphy et al., 2019), or could result from a spatial contraction of summer floods in the future, which have historically resulted from short-duration, high intensity storms.

We appreciate the reviewer's suggestion and agree that it is more appropriate. The change suggested is applied.

L 212: It's unclear what the term "methods" is in reference to. Is it the method of climate modelling or is this meant to mean differences in flood responses to different storm types?

Methods should be "models". The sentence now reads "Thus future work could build on Kay (2022) and look more specifically at differences between RCMs and CPMs which could explain these patterns."

Figure 4: The left figure is superfluous. Please remove this.

Agreed and removed.

L227-230: Describing the changes in reference to changes in return periods is unconventional and is probably due to the way the figure axes are configured (see comment on Figure 5). Please change this description to one that describes the shift in event duration with respect to frequency.

We adjust this sentence to read "… somewhat correlated, in that the events with the highest return periods are very unlikely to be of short duration.

Figure 5: is there actually a heavier tail in the future in SON? (L223) Also, the return period must be plotted on the x axis, as the duration of the event is dependent on the return period not the other way around. This isn't just a formatting choice – the reader is forced to attempt to transpose figure 5 as the convention in referring to tails is the distribution along the horizontal (with a few exceptions). In addition, there isn't a good reason to present Figures 5 and 6 in a way that is disjointed. The reason given in response to the previous reviewer of duration having a smaller range than return period (and likewise in Figure 6 the return period having a smaller range than area) is disputable: one could easily represent the return period in log10 years and have an effective scale of 0-4 (just making a point – I'm not suggesting that return periods be presented in this way).

Figure 5 has been flipped to be consistent with the other figures in this paper, addressing the problem you present.

Figure 7: labels a, b, c, d are missing from Figure 7 (b is referenced in the text). The caption needs to describe what is in each of the four figures (i.e. different time slices and different measures). This figure is also inconsistent with previous figures that have shown the baseline time slice on the top row and the future on the bottom (this can be fixed and the legends could simply be placed horizontally under the respective columns). Using a different color scheme for each metric would aid in interpreting the different scale of results.

Fixed as per suggestions.

Figure 8: as per Figure 7 regarding layout and caption. Only one legend is needed.

Fixed.

Minor comments

Grammar and style

L66: “of the present work”

Removed as relevant to bias correction work elsewhere

L175: “shown” instead of “show”

Fixed

L231-233: missing “there”; “matched” not “matches”; “dynamics” is not an appropriate adjective here. I think what is meant is: but there is variability amongst the ensemble members in the relationship between event duration and frequency.

Fixed: replaced “dynamics” with “patterns”.

L237: “is not surprising”: please replace this with objective language e.g. statistically plausible

Fixed

Referencing: Following on from a previous reviewer’s comment, please check the accuracy of all references. One example is that of Tawn et al. 2018: it is referenced as 2019 in the text and the list of authors in the reference list is incorrect. The dates for Towe are also reference incorrectly in text. There may be others, so it may be prudent to check the referencing system.

This reference has been fixed, and re-proofread. There are two references here which have been confused, and so this has been corrected: Tawn et al. (Spatial Statistics, 2018) and Towe et al., (J AGRIC BIOL ENVIR S, 2019).

Reviewer 2

General

As stated previously I do think the strength of the paper is its use of a climate model ensemble to simulate areal flood events, where the joint interaction between the factors that cause floods over a range of temporal and spatial scales is implicitly accommodated by the use of a gridded daily continuous simulation model. I have gone through the various comments and responses, and overall I appreciate and agree with the changes made by the authors.

I appreciate that the authors have undertaken a computationally demanding set of simulations – the data sets and modelling framework are impressive – but I regret that I am left with concerns about two major points which I raised previously, namely 1) the need to compare baseline modelling with observations, and 2) the defensibility of the probability of exceedance estimates. The authors’ response to the former point focussed on the problems with bias correction without addressing the more fundamental concern around the need to demonstrate how well the modelled baseline frequency curves of areal rainfalls or floods conform to observations. The response to the latter point (the method of calculating probabilities of exceedance) does not provide additional confidence in the validity of the results and some reconciliation of the different estimates is needed.

This is discussed in more depth for Reviewer 1’s very similar point. In brief, no gridded flow observations exist for the UK, but we compare the results to “observation-driven simulations” as used in Kay (2022). With regards to bias correction, we include a brief discussion where it was shown that, nationally, bias correction at locations with stations systematically underestimated return periods across the country compared to those observed, and without bias correction this systematic underestimation was lessened. Thirdly, the distribution has been swapped with a standard Generalised Pareto distribution using a simple scaling factor to convert from a per-exceedance to an annual probability.

I provide more commentary on these two points below.

Baseline Evaluation

I agree with the authors' caveats about the dangers of bias correction, but my main point in this regard was the need to provide evidence that the frequency distribution of areal event extremes derived from the UKCP18 data compare reasonably well with observations. My earlier comments expand on this point a little, but as far as I can tell the authors' justification for not examining such evidence is because they adopted flow thresholds to yield a specified number of flood exceedances over a given period (and over a given area). This approach does not demonstrate how well the results generated over the baseline period relate to real-world conditions, it is merely a device to extract the number of events relevant to the exceedances of interest. What is missing here is a comparison of the selected thresholds with what has been observed over some suitable baseline period.

In other words, the selection of thresholds to yield a defined number of flood occurrences provides no information on how well the magnitude-frequency relationship of the flood regime is preserved (ie how well the cumulative density function governing the extremes matches reality), and this is a particular problem as the modelled floods are fitted to a Pareto distribution and the results are reported on in terms of absolute shifts in return periods (ie in terms of a shift in the magnitude-frequency relationship).

The results derived by fitting a Pareto distribution to the modelled events are very likely to be impacted by various forms of bias, and varying the thresholds to achieve the required number of exceedances does not avoid or obviate the need to undertake bias correction, it just ignores the problem. The degree of difference between modelled and observed flood frequency curves for some representative locations would add considerable weight to the conclusions; without demonstrating this, the conclusions would need to state clearly that the adopted methodology yields results that have not been evaluated against observations over the baseline period, and thus are rather more speculative than currently framed.

See our discussion on using SIMOBS in our comments to reviewer 1 for more detail. There is a lot of new content throughout.

Probabilities of Exceedance

I also think we should be troubled by the difference in results obtained using the two procedures to estimate the probabilities of exceedance (ie the one based on deriving annualised exceedances from a fitted GPA distribution with an assumed spatial dependence, and the method as described in the paper).

The authors have stated that they wish to retain the method as outlined in the paper as it is being used in another manuscript and they wish to "minimise confusion". While I appreciate the convenience of this decision, if the method cannot be shown to be correct then I see little point in being consistent.

The problem is that we have two methods (applied with different levels of rigour) that have two markedly different risk implications: I suggest that the subsequent check made with the assumed degree of spatial dependence yields results that are consistent with expectations, whereas the results presented in the paper are not. I think this a problematic outcome.

Accordingly, I think additional investigation is required to reconcile these different probability of exceedance estimates. The differences in results are too great and too surprising to ignore, and some diagnostic checks should be devised to check whether the inconsistencies are due to computational errors or unsupportable assumptions in one or other of the approaches.

Rory Nathan University of Melbourne

We thank the reviewer for this considered response. To reduce the complexity of both the computation and the discussion, and since we only care about threshold exceedances, we restrict this paper to a standard Generalised Pareto distribution as suggested in the previous review. Since the probabilities are still "per exceedance", we convert to annual probabilities using a simple scaling factor (# events extracted ÷ #

events per year). This reduces the number of events with a return period exceeding 1000 years to a more plausible level.

Reviewer 3

The paper investigates the change to wide-spread flooding in terms of number of occurrences duration and extend between current and future climate using an ensemble regional climate model. This is a very relevant study showing novel results. However, revisions are required as per the comments below.

Main comments

The novelty and aim of the paper is not clear after reading the introduction. In the introduction the relevance of the work is clearly discussed, however no clear research gaps are mentioned. Also, at the end of the introduction (L46-50) a short summery of the methods rather than the aim of the paper is provided.

Some clarifications are needed in the data and methods sections (see specific points below). Especially, the statement that bias correction of precipitation from climate models is not required when focusing on extremes needs explanation. Furthermore, I think if the sensitivity of the assumptions (event magnitude threshold, extent threshold, and maximum event duration) in section 3.2 should be discussed separately from the methods as the current structure makes it hard to follow the methodology. And the manuscript would really benefit from a discussion on how these assumptions affect the final results (rather than the number of events for a single ensemble).

The conclusion section also contains discussion of the results as well as some limitations. For clarity, I recommend splitting the conclusions and discussion by making a separate section on limitations and moving the discussion of the results in relation to other papers to the "results and discussion" section.

Firstly, we add at line 50 "Often flooding is considered on a site-by-site or regionally summarised fashion, particularly when looking into projections of the future. This paper hopes to show the benefits of considering widespread flooding events over a large area using gridded, rather than catchment-based hydrological modelling to expand our knowledge of the extent of possible flooding events in the UK. Comparing UKCP18-driven model runs to those driven by observed rainfall and temperature will give confidence to the use of these event sets in future analysis." To the introduction to highlight our aims and the novelty of the work in the rest of the paper.

The discussion of bias correction is now greatly lengthened, see out comments to Reviewer 1 for more detail.

The discussion and conclusion sections have been restructured to remove new discussion points from the conclusion, and move limitations to the discussion section as well. Discussion points which do not specifically count as results have also been moved to the discussion section.

Specific comments

L12: "…, allowing events to last up to 14 days" seems to suggest that the 14 days are a consequence of the event definition given before, but I don't see how this is the case. Can you clarify?

Changed to "with a maximum duration of 14 days" for clarity.

L36: I suggest to leave out "driving" in "driven by large ensembles of driving data from climate models" as it seems double?

Removed the word "driving"

L37: Can you provide more information about what "predominantly stochastic event-based models are"?

We replace this with "…, or through Monte Carlo methods simulating boundary conditions to feed into an event based model (PQRUT) (Filipova et al., 2019)."

L42: Replace "focusing on the United states" by e.g. "for the United States", as using "focusing" twice in this sentence is confusing

Replaced with "for the United States"

L54: This is the first time UKCP18 is explained while it has already been used in the abstract and introduction. Please explain at the first use.

All abbreviations are now explained at first usage.

L66: The authors seem to state that bias correction of precipitation from climate models is not required when focusing on extremes? I would disagree with this and would like to see more better argued why no bias correction is required.

As discussed above for Reviewer 1, a small example is given showing that bias correction leads to a systematic underestimation of the 2- and 50-year events observed at stations.

Section 2.1: Why was RCP8.5 selected? It would be good to mention this choice also in the abstract.

RCP8.5 was the only scenario available from the UKCP18 gridded rainfall datasets at the required resolution – there was no choice available to be made. We add "- the only available scenario in the 12km grids (Riahi et al., 2011)" to line 59.

Section 3.1 It would be good to add which version of the model is used. Whether the model has been calibrated for this study or a previously calibrated version has been used.

We add the following to line 80: "This paper uses the same version of Grid-to-Grid as used in Kay et al., 2018 and Kay 2021, since it uses the same driving data."

L105: "this was considered equivalent ….". This subsentence is confusing to me. Please consider leaving it out or clarify its meaning.

This subsentence is replaced with "(denoted 'inundated')".

Table 1: "PoE" is not explained. Furthermore, it would be clearer if consistent naming of thresholds were used in the "exceedances" column and in the text (e.g. POT2 or 2/yr). Also note the typo in "exceedances".

Another proofread has been undertaken, and all abbreviations are spelled out on first occurrence.

L119: I don't understand why based on minimum-threshold, flood extents that are smaller than this threshold are "retained". I would expect these are excluded. Could you please clarify?

Sentence changed to "The 0.1% inundation coverage was selected to ensure that small, very extreme events were not excluded."

L154: Why were 60 peaks selected and how does this relate to earlier event magnitude threshold? Also, from which ensemble or for all ensembles? Please clarify.

60 peaks matched an average of two events per year in a 30-year timeslice. Each ensemble member was fitted separately. We have added "the top 60 independent peaks *in each ensemble member and timeslice* were found…".

L159: How is the "daily exceedance probability" calculated from 60 peaks? I'm used to converting these probabilities to annual exceedance probabilities directly based on the average number of peak events per year, which you seem to refer to as an "alternative approach" (L171).

As discussed for Reviewer 1, in this paper, we have switched to just using a standard Generalised Pareto distribution, rather than the previously stated mixed distribution, and as you have mentioned, swapped to converting to annual probabilities using a simple scaling factor (# events extracted ÷ # events per year).

L174: Could you provide more context to the sentence “which might potentially align with discussion of the frequency of 100-year events in the UK”? What is this discussion about?

This sentence has been removed to improve clarity.

Section 4: In Figure 3, 5, and 6 event return periods are shown (if I understand correctly). It is however unclear to me how these are calculated? As show in Figure 2 the return period varies spatially, and it is not clear how a single return period is calculated.

The sentence “In the rest of this section, return periods reported in the text and figures are the maximum return period observed (across space and time) within a single event.” has been added to line 186.

L234: I assume AMAX is “annual maxima”? This has not been explained before.

All abbreviations are spelled out on first occurrence.

L237: What do you mean by “change in flow”? Is that change in peak event magnitude?

Changed “flow” to “event peak flow magnitude”.

L247: Can you explain what the value of 120 km is based on? In the figure it seems there are still points whit significant asymptotically dependence up to 250 km.

In this sentence, we add “… have a limit at most location pairs of around 120km…”

L271: I would be very useful to understand how sensitive the change in number of events is to the selection of thresholds in section 3.2. And how significant it is given the differences between ensemble members.

We add this sentence to line 114. “ Very similar patterns of events extracted (not different at a statistically significant level) were observed for all of the ensemble members.”

L299: Can you clarify what you mean by surface water flooding as opposed to fluvial flooding?

Surface water flooding refers to flooding caused by means other than the overtopping of a river or water body. For example, high-intensity rain storms in paved urban areas can result in surface-water flooding if drainage is insufficient.

Editor Comments

The overall comments from the reviewers highlighted to us the need to go back and do further work to justify the use of a) the modelled UKCP18 data and b) not to use bias correction. We also reduced the complexity of the probability distribution used to a simple GPa distribution with a basic scaling factor to get annual probabilities. This leads to a better link up to other work in this field. This all led to two new sections of work in the paper (explained in our comments to Reviewer 1) and we feel it has benefited the paper over all. We hope that this addresses all the remaining concerns.