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.

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.

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