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.

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.

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.

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.

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.

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.

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?

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.

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.

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.

Rory Nathan
University of Melbourne