General Comments

The overall concept of this paper is neatly done: a 12-member ensemble of baseline and future climate (12 km resolution) is input to a grid-based hydrological model (1 km resolution) to characterise the impact of climate change on flood events. The strength of the paper is in its focus on 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. All inferences about changes to flood risk are made using 30-year sequences of daily floods, as derived from the 12-member ensemble of climate projections. Differentiating impacts by the areal extent and duration of floods of varying severity is novel, as is the exploration of possible changes in their spatial dependency.

There are, however, some aspects to this work which are potentially problematic, and these need to be addressed by further explanation and/or revision.

Specific Comments

The key issues that I am struggling with are as follows:

- It is difficult for a dynamically downscaled rainfall products to reproduce rainfall quantiles over the temporal and spatial (meso-) scales relevant to catchment flooding, and I was surprised to read (lines 62-63) that "due to the focus on … extremes rather than the whole regime in general" that no bias correction was applied. Bias-correcting projected extremes is as important, if not more important, than a central tendency measure. The rainfall-based simulation of floods is critically dependent on the correct representation of the frequency distribution of areal rainfalls, and I think it important to provide evidence that the frequency of areal rainfall extremes derived from the UKCP18 data compare reasonably well with observations. To this end, providing evidence that distributions fitted to n-day maxima extracted from UKCP18 (preferably for a range of areal extents relevant to the adopted spatial limits) are reasonably consistent with those based on observational data. I searched for any such evaluations in the Met Office documents (the citations provided for these need to be improved and corrected in the manuscript) but I could not find anything specifically relevant to the rainfall behaviour of most interest.
    - There are many issues and assumptions inherent in the bias-correction process, including the assumption that the same 'biases' seen in baseline climate model data are also present in data for future periods, concerns that correction can alter the spatio-temporal consistency of individual variables or break important physical relationships between variables, and the fact that typically-applied daily rainfall corrections can fail for multi-day rainfall totals (e.g. Ehret et al. 2012, Addor and Seibert 2014). The application of bias-correction can even introduce artefacts into the 'corrected' data (Maraun et al. 2017). Attempts to 'correct' rainfall extremes are especially difficult, as by their nature they have limited occurrence in observation-based datasets and are also strongly affected by natural climate variability (e.g. the well-known presence of prolonged flood-rich and flood-poor periods), whereas ensembles of climate model data will not necessarily present the same 'states' of natural variability through time. Thus application of bias-correction, rather than reducing uncertainty, represents a considerable source of uncertainty in itself (e.g. Lafon et al. 2013, Ehret et al. 2012). In this application we instead chose to use only the raw climate model data, to maintain the spatio-temporal properties of precipitation, temperature and potential evaporation imposed by the dynamic downscaling of the RCM. We then determine what constitutes an "extreme" level of flow by selecting a threshold based completely off the climate model runs, not observations and hence any bias in threshold selection is matched by bias in the events. The upshot of this is that the key features and results are not impacted by the bias.

- On the basis of the information provided it is difficult to be comfortable with the reported probabilities of exceedance (PoE). In concept the approach of adopting a merged CDF on the basis of empirical and fitted distributions is fine, my difficulty is with the inferred annual PoEs. I suspect that there is a problem with the way that the Poisson approximation is applied, and I suggest that the authors compare (or replace) their analysis with the more straightforward approach based on fitting the GPA distribution to the POT2 series, where the annual quantiles are obtained by the simple expedient of factoring the exceedance probabilities by N/M, where N is the number of years in the record and M is the number of maxima extracted. The key reason for my discomfort with the PoEs reported is the severity of the identified events. For example, in Figure 2 it appears that 3 (possibly 4?) events with return periods of 1000 years have been observed in a single 30 year sequence. I appreciate the need to consider the influence of spatial dependency and the trading space for time issues here, but still, this number of extreme events is higher than expected (and higher than I suspect would be extrapolated by Tawn et al, 2019). A crude estimate of the likelihood of this could be obtained by estimating the notional number of largely independent catchments across the UK. If we adopt a spatial dependence limit of 120km (from line 220 in the paper) then the notional upper limit of the spatial extent of an event might be around 45000 km^2, which yields around 5 or so independent catchments (or "trials") in each year. Given that the likelihood of a 1 in 1000 event occurring in a 30-year period is 0.029 (from the Binomial distribution), then there is about a 13% chance you would see a single 1000-year event in one of the five independent catchments somewhere across the UK in a 30-year period. However, we would actually need to have around 50 independent catchments in the UK to see three 1000-year events occurring in a 30-year period with any likelihood, and this corresponds to an asymptotic dependence limit of only around 40km, which is very low given the information presented in Figure 7. The number of exceedances shown in Figure 5 is larger again, but this may be due to how the ensemble members are combined (discussed in the next point).
    - This is a really interesting point, and quite an insightful way of estimating the number of very extreme events within a given period. The merged CDF is required for the copula method to be applicable, however, the empirical component of the distribution is not actually used in the figures since the threshold for using the GPa exactly corresponds to our threshold for delineating the event extents. A preliminary investigation suggests that your alternative greatly reduces the return periods of the most extreme events, reducing most of the >1000 year return periods to under 1000 years. However, we have a second paper in publication building on this work, and the authors feel that a consistent presentation of return periods across the two papers would minimize confusion. As we feel this is an important point to make, we will include a paragraph at line 156 outlining the alternative approach, with an example flood frequency curve to highlight the differences.
- If my understanding is correct (lines 175-177), the 12-member ensemble from UKCP18 has been lumped together and used in the preparation of the results as summarised in Figures 3 to 7. I think this approach confounds the absolute interpretation of the reported frequencies and return periods, and I suggest that it would be more useful to treat each ensemble member as a source of aleatory uncertainty over a 30-year period. Thus, rather than reporting, say, that there are 17 events larger than 1000-year event in DJF (Fig 5) under baseline conditions, it would be more useful to report on the average (or median) frequency/quantile across the 12-member ensemble, where the highest and lowest ensemble member provides an indication of the upper and lower bounds of the sampling uncertainty in each 30-year period.
    - This is a good point. Aside from Figure 3 which is already split by ensemble member, we can easily include uncertainty bounds on Figure 4, and include some measure of variance in Figures 5, 6 and 7 through adjusting transparency (alpha), where low variance is shown by a stronger colour, and high variance by fainter colours. Including upper and lower bounds in addition would result in a lot of extra figures, or much more complex ones, at a cost to

readability. At line 177 we replace the sentence “In the rest of this section, the event sets …” is replaced by “In the rest of this section, ensemble members are treated as separate sources of equal weighting. Variance between ensemble members is indicated in figures by brightness, and the colour indicated the median value of the respective statistic amongst the ensemble members.

- Lastly, no discussion is provided on how the asymptotic independence metric varies with distance (lower panel, Figure 7). I think the metrics used by Coles to explore asymptotic behaviour would benefit from additional explanation here as they are not intuitively obvious; specifically, the way in which the independence metric is defined is easily misinterpreted and without explanation it appears odd that the degree of independence is decreasing with increasing distance, which is exactly the opposite of what one would expect (and as shown in the dependency metric in the upper two panels of Figure 7, which is consistent with intuition).
    - This is a reasonable point to make. On the one hand, we do not wish to just repeat Coles, however we agree that intuition may be misleading. We edit the text at line 136 to the following: “…are calculated between pairs of points. For two points i and j,

$$\chi_{i,j} = \lim_{x \to \infty} P[Q_i > x \,|Q_j > x]$$

      If C*(u,v) = 1 – u – v + C(u,v) for a copula C between two points i and j, then

$$\bar{\chi} = \lim_{u \to 1} \frac{2\log(1-u)}{\log(C^*(u,u))}$$

      χ describes the level of asymptotic dependence, if χ = 0 then the variables are asymptotically independent, otherwise they are asymptotically dependent. In the asymptotically independent case, $\bar{\chi}$ describes the dependence for large but not asymptotic values of flow. In the asymptotically dependent case, $\bar{\chi} = 1$.”

    - To comment on both panels of Figure 7, we change the sentence on line 220 to: “The figure suggests that asymptotic dependence decreases as distance increases. In the asymptotically independent case (Figure 7b), we see a similar pattern in dependence for large values of flow, with high dependence at short distances, even if they are independent in the limit.”

Additional References

Addor N and Seibert J (2014). Bias correction for hydrological impact studies – beyond the daily perspective. Hydrol. Process. 28, 4823–4828.

Ehret U, Zehe E et al. (2012). HESS Opinions "Should we apply bias correction to global and regional climate model data?" Hydrol Earth Syst Sci, 16, 3391–3404.

Lafon T, Dadson S et al. (2013). Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. Int. J. Climatol. 33: 1367–1381

Maraun D, Shephard TG et al. (2017). Towards process-informed bias correction of climate change simulations. Nat Clim Change, 7, 764–773.