Response to interactive comment of Anonymous Referee #2

We would like to thank Anonymous Referee #2 for their constructive feedback. The reviewer indicated that the manuscript is promising and interesting, and well written. With the suggested feedback, we expect to further improve the manuscript.

Generally, I like how the authors approach the problem and believe the article is clearly written and to- the-point. It is relatively short, but concise. Nevertheless, there are several issues that the authors may need to address.

First, I am not sure if the parameter sampling strategy is sufficient. A sample size of 100 parameters is, in my view, extremely low. I like how the authors use a K-S-test to assess whether the sampled distribution differs from a benchmark set, and believe also that this could be a good approach to determine the appropriate number of samples. However, the benchmark sample size is also just 500 samples, which is also still relatively low. With eleven parameters, this means that the sampling density (defined as N^(1/p), with p the number of dimensions and N the sample size), is just around 1.76. In other words, on average, there are less than two samples per parameter. I think this sample size should be increased to at least a couple of thousand, then the KS-test makes more sense and can be used to select a lower, proper number of samples for the rest of the analysis. Of course, I fully understand that there will be a computational burden to it, but the authors could do this also for a shorter time period as the 34 years used now in order to save resources.

We agree that the parameter sampling is rather coarse, indeed because of computational constraints. Testing for shorter time periods however, has the disadvantage that we then cannot test the effect of parameters on the kind of events we are interested in (extreme events with long return periods, 34 years is already relatively short for that). Furthermore, we would like to emphasize that we used a Latin Hypercube Sampling Strategy, this means that for a sample size of 100, each parameter has 100 different values because the parameters are all sampled at the same time (this can be done under the assumption that the parameters are independent).

Based on the feedback of the reviewer, we have increased our benchmark sample size to 5000 (this used to be 500). The results are shown in Figure 1. We still observe that the D-statistic starts to stabilize at around 100 parameter samples, therefore we do think we can safely assume that a sample size of 100 is a reasonable size to capture variability introduced by parameters. This number seems smaller than found in many other studies, and probably relates to our variable of interest - only the maximum and minimum discharge.

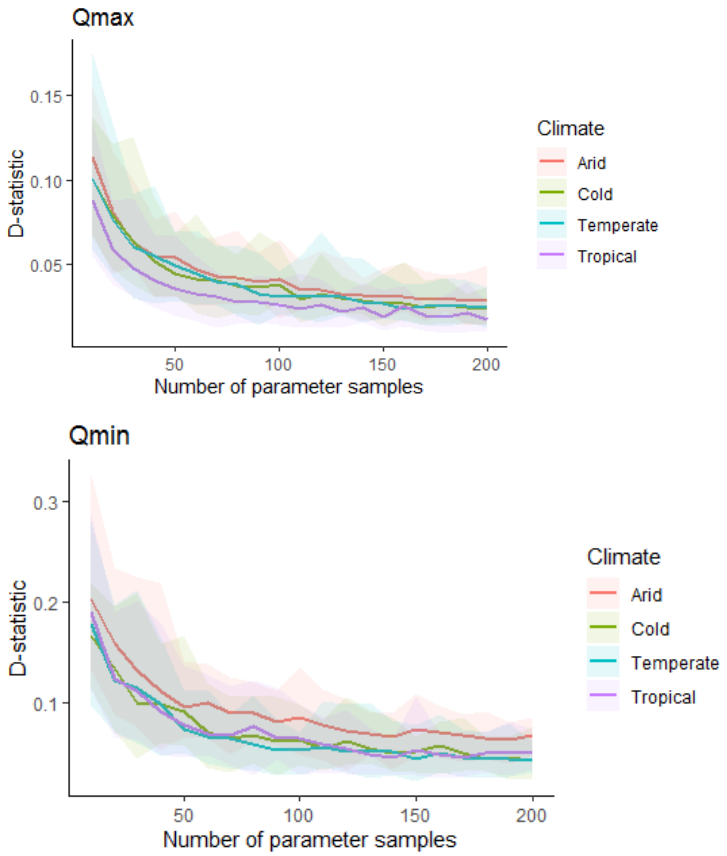

Figure 1. Overview of the D-statistic for Qmax (upper panel) and Qmin (lower panel) across the four climates. De D-statistic is obtained benchmarked against a sample of 5000 parameter sets.

The authors are also quite critical on their own results regarding the low flow events, which is a very good thing in itself. However, if there are indeed so many numerical artefacts here, and we can not fully trust the results, it may just be better to completely leave this analysis out and focus on the high flow analysis.

We have considered focussing only on high flows based on the results, but in the end made a deliberate choice to include the low flow results as well, to overcome the so-called “publication bias” where only positive results are published. We hope to avoid people repeating the same study for low flows and finding the same problems, because we decided not to publish or include the problems arising when evaluating low flows. Furthermore, the negative results on the low flow events may guide further research efforts into improving the modelling of such flows.

I also wonder how much the cell-based approach matters. Especially regarding floods, the size of the catchments matters, as the flood-wave will be routed through the river-network. There was no routing model included, so how much will this make a difference

in the results? Or, in other words, are the cell sizes small enough to ignore the routing effects?

Indeed when applied to a specific catchment, the catchment size and the temporal resolution will determine whether routing can be ignored or not. Routing will delay the peak and decrease the peak height. We did not consider routing because we apply a synthetic experiment and therefore the routing parameters cannot be calibrated on a catchment outlet measurement station. The effect of the routing parameters on the peak are known (delay and attenuation) and consistent among the different model structures if the same routing procedure is applied: the routing has no effect on the generated runoff itself. In this way, we keep the comparison clean. It is, however, a valid point raised by the reviewer and we will add it to the discussion.

Lastly, it is not fully clear to me how the analysis on the timing of the extreme events works. Why do the resulting bar charts in Figure 7 have a varying number of events on the x-axis? Do these correspond with different parameters, model structures or different return periods?

The timing analysis is indeed rather complex to explain, we will try to further improve the description of the procedure. To explain the numbers on the x-axis: Since we evaluate the timing of events with a 500-year return period and we have a simulation period of 2000 years, each simulation will have 4 of these extreme events. If all the different simulations (with combinations of different parameters and different model structures) agreed upon the timing of this extreme event, indeed only 4 events would be identified in total, and the x-axis would go to a max of 4 with 4 fully filled stacked bar charts (indicated as the “theoretical max”). The number on the x-axis indicates the number of extreme events with a different timing. So, if the x-axis goes up to 20, it means that across all the simulations, 20 different 500-yr return period events with a different timing can be found. The higher the number on the x-axis, the more variation there is among the different simulations in the timing of 500yr-return period events. The height of the bar chart indicates how many simulations identified a particular event. In the temperate climate, for instance, 1 event is identified by all simulations because it has a fully coloured bar chart. However, there is large disagreement about the timing of the other 3 events given that 38 events with different timing were identified.

To conclude, the manuscript is very promising and interesting. I really like the methodology, and think the article is well written. I hope the authors find my comments useful and I look forward to an improved version of the manuscript.

Thank you!

Minor comments

Suggestions for textual corrections and textual additions for clarification will all be implemented.

P7.L151-152. The Kolmogorov-Smirnov...is significant. → This is a bit unclear, but I think you mean that you asses whether an empirical distribution based on one parameter sample is statistically different from another empirical distribution based on another parameter sample, correct?

That is correct, we will reformulate.

P9.L191-192. I assume you also repeat this for each parameter set for each model structure, correct?

Step b is the evaluation at the parameter set level, and step c is for each model structure. In the end, indeed, both parameter set and model structure are accounted for in this analysis. In response to this comment and the comment on Figure 7, we will further elaborate on this approach.

P10.L203-204. Why just four of them? I think you should also show the others, at least in the Supplementary Material.

We selected these four models to demonstrate and explain the analysis (that's the goal of this figure) - we thought that including all the models would make it less clear. The results for all the models are shown in Figure 6. We will test whether we can include all models in Figure 5 while still keeping things clear, e.g. by showing all other models in grey.

P14.L320. Which are .. stability → why not check this with the developers?

This is a good suggestion, we will inquire with the developer (Martyn Clark, and Nans Addor which is currently maintaining the FUSE code).

P14.L.325. Our results...these events. → Can you elaborate and clarify how I can see this? I am not sure how the see this from the results as shown so far.

This is indeed not indicated very clearly in the previous parts of the results section. We will evaluate our statement and indicate references to this conclusion in the result section where applicable/appropriate.

Fig3. The unit of probability seems strange to me. Is that correct? Please also add a legend.

It would be more correct indeed if the y-axis would be labelled as probability density rather than probability. Since the area under a probability density plot needs to be equal to one, the units of probability density are the reciprocal of the units of the x-axis, and thus correctly indicated in the plot. A legend will be added, thanks for the good suggestions.