Review of ***The impact of hydrological model structure on the simulation of extreme runoff events*** by Van Kempen et al.

The manuscript of Van Kempen et al. deals with the influence of model structures on the magnitude and timing of extreme events. To do so, the FUSE framework was used with ten model structures and 100 parameter sets. The models were applied for four different climate zones and forced with a simulated timeseries of 2000 years. The authors show that alterations in percolation and evaporation affect mostly the magnitude of high flow events, especially for the cold and temperate climate zones. For low flows, especially the lower and upper formulation mattered. Generally, the model structural uncertainty was found to be higher for the low flow situations. In the arid and tropical climate zones, almost all model simulations agreed on the timing of the events, which showed a reduced influence of the model structure.

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.

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.

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?

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?

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.

Minor comments

P2L28. on a statistical models → on a statistical model

P2.L34. Statistical model → statistical models?
P4.L96. I am getting a bit confused here, are you setting up the models for four selected grid cells? Can you provide some details?
P5.L126. Selected structures → selected model structures
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?
P9.L189. A high value in the red row → Maybe define also what this value is in the text.
P9.L191-192. I assume you also repeat this for each parameter set for each model structure, correct?
P10.L203-204. Why just four of them? I think you should also show the others, at least in the Supplementary Material.
P10.L207 as displayed in Figure 6 → Please help the reader a bit, how do I see this exactly?
P10.L211-214. This section...Figure 3d. → This sounds as if this text was originally below the header of section 3.1 and later moved here.
P10.L221. Has least impact → has the least impact
P12.L255. Soil moisture → The upper layer soil moisture is actually quite critical in extreme events, as with saturated soils more overland flow occurs.
P14.L320. Which are .. stability → why not check this with the developers?
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.

Fig3. The unit of probability seems strange to me. Is that correct? Please also add a legend.
Fig6. Please add what you mean with the lower limits in the caption, and an x-label with the return periods.