Replies to reviewers – Submission titled "Historical and future changes in global flood magnitude – evidence from a model-observation investigation" [HESS-2019-388]

Replies to Reviewer #1

We would like to thank Reviewer #1 for the constructive comments which will help to improve the quality of our manuscript. Each of the reviewer's comments will be quoted in *blue, italic text*, followed by our reply formatted in normal text.

General comments

Not content related: the authors list reads like a "manel". Sad to see so little inclusion of other genders than male. Maybe something to think about for future studies?

We thanks the reviewer for their note about the broader diversity in the author panel in such global scale initiative. We will aim for improved inclusion of gender diversity in future investigations.

I do miss the overarching research motivation of this study. Even though the introduction contains a good amount of background literature and describes the three research objectives set, it does not become clear what the research gap is that the present study intends to fill. Or in short, how are the objectives derived and what is the (societal) relevance of the study? Please ensure this is clearly and concisely explained in the revised version of the manuscript.

Thank you for your note that the research motivation should be presented more prominently. We will revise the introduction and consider revisions focusing on: (i) highlighting the contribution of this study to address existing research gap in the field, (ii) including the motivation from the perspective of global hydrological model users (e.g. decision makers).

The models used in the study differ in quite some characteristics/schematization. An overview of these differences would be very useful (also possible as supplement). Besides, none of the deviations between results obtained with different GHMs is related to those different characteristics/schematization. I would guess that they do play a big role in explaining the obtained results and thus I recommend extending the manuscript with such an analysis (or a clear and convincing statement why not).

We appreciate the reviewer's suggestion about exploring the role of model conceptualization in simulating trends in floods. We will explore the possibility of having a supplementary table outlining the differences among GHMs in the revision.

We note, however, that a detailed documentation of differences in model characteristics does not necessarily enable evidence of why the models produce different outputs. From our understanding, previous studies have related the impacts on peak flows of some specific processes such as routing scheme or reservoir algorithms (Zhao et al., 2017;Masaki et al., 2017), but it does not mean that a similar impact is presented in simulated trends (also highlighted in our manuscript at Line 85). This is also a motivation of this study, as we want to compare the trends of a high flow indicator simulated by different GHMs.

To nevertheless incorporate the reviewer's comment, we will make the main objective of this study (i.e. a comparison of model capacity in simulating trends in a flood indicator and an assessment of uncertainty of projected changes in floods) more prominent in the introduction. We will also mention the reviewer's suggestion (i.e. to assess the impact of model schematization on changes in flood hazard) in the conclusion as a potential research direction.

The manuscript reads bit lengthy sometimes. To some extent, this is the result of reporting a lot of numbers, which are partially also provided within tables. My recommendation would be to let the tables and figures speak for themselves and to check which reported numbers can be neglected as they are not directly needed for understanding the methodology or results. Plus a check whether more concise wording could be used.

The manuscript will be carefully revisited to focus more on the key findings, reduce any redundant information and present the results in a more concise manner. We will focus especially on Section 3.1 where we will remove any descriptions that have already available in the tables or figures.

This study is very much focussed on data and their analysis. Extending the implication of the findings of the study to the societal dimension would greatly benefit the manuscript to make the results more tangible and applicable.

We thank the reviewer for the suggestions on extending the study findings to societal dimension. We note that the key motivation of this study is to explore the level of consistency of trends detected from streamflow observations and model simulations, which is currently underrepresented in the literature. As a result, we would like to focus on this research pathway and will refine the introduction to highlight this objective better.

I would very much welcome it if at least the data pairs for the observation stations are provided via a supplement. This would be in the spirit of FAIR hydrologic modelling, thus increasing reproducibility of your findings.

We will upload the observed and modelled trends at each station as a supplementary (csv files) together with the revised manuscript.

Specific comments

• P1/L29+L30: An agreement of 12-25 % can hardly be named "moderate agreement", can it?

The abstract will be revised to ensure appropriate terminologies are used, potentially by changing "moderate" to "low-to-moderate".

• P1/L31 "significant differences": please specify what kind of differences you refer to.

We will clarify that the characteristics of trends (trend mean, trend standard deviation) simulated by GHMs forced with historical climate is significantly different to that simulated by GHMs forced by bias corrected climate model output.

• P2/L52 "specific regions": specific regions such as? Please specify.

We will clarify that this the statement may only applicable where rainfall plays the dominant role in flood occurrence.

• P2/L53 "recent evidence": what evidence? Please provide name, source, etc.

The following sentences (Lines 54-56) in fact have extended our discussion and provided some evidence for this statement. We noted that this may be unclear and will revisit this paragraph to ensure the statement is justified.

• P2/L52-L66: what about the role humans play in changing flood hazard? Please add this dimension to the paragraph.

We will add the impact of human activities to changes in flood hazard at the end of this paragraph.

• P3/L77 "factorial experiments": what are you referring to with this term? Please explain or use more common terminology.

"Factorial experiments" indicate studies analysing the effect of different factors (e.g. land use change) on the response variable (e.g. changes in floods), as well as the effects of interactions among the factors on the response variable. In the context of hydrological modelling, the impact of atmospheric forcing, land use change and other drivers of change on streamflow trends could be "turn on/off" to provide a full "factorial experiment design". We will rewrite this statement to improve readability in the revision.

• P3/L87-L100: The description of the research objectives could profit from using bullet points

We thank the reviewer for their suggestion. We will consider using bullet points to separate the research objectives.

• P4/L105-L108: why are these models used? Why are some others available within ISIMIP not used? Please clarify.

There is no model selection in this study. We actually used all GHMs that have provided discharge data within Phase 2a and 2b simulations at the time this study was initiated (June 2018). In the revision, we will highlight this fact in a transparent manner.

• P4/L122-L126: As mentioned in the general comments, a (technical) description of the models is needed. A particular focus should be on how the models changed between ISIMIP2a and ISIMIP2b and whether those changes may have influence results (or not). Possible changes in e.g. functionality, spatial resolution, etc. may have had a great impact on results and thus affecting the comparison performed in your study. I thus strongly disagree that checking this is outside the context of the study.

Similar to our response to the general comment from the reviewer, a precise conclusion about the impact of changes in models (e.g. functionality, spatial resolution) on trends in floods should be based on a full multi-model experiment (i.e. to compare trends simulated by different versions of the same GHM), which is unfortunately not readily available. Although we are aware that changes and bug-fixes done in MPI-HM affect only the human impact simulations (and the influence is insignificantly), it is not straightforward to generalize this conclusion. We will aim for some extended discussion, but would like to keep our statement as-is (i.e. checking the affects is outside the context of the study).

We also note that the issues raised here and in the earlier comment show the need for the next step of model inter-comparisons which should focus on diagnosing the reasons for differences across models. We have mentioned about this need in our manuscript (Lines 542-547) and will consider make this call more prominent in the revision.

• P6/L144-L147: what was the reason to not only use the un-routed runoff for all catchments? Wouldn't this increase comparability between results as it removes (unnecessary) transformation of results and units?

The reason is that for large catchments, observed discharge and unrouted runoff are not comparable. In some very large basins, it takes one to three months for upstream runoff to reach river mouth through the channels (and be measured as discharge here by some observing gauges). The same magnitude of basin total runoff, depending on its spatial distribution (i.e., evenly distributed versus concentrated in the downstream), could generate rather different discharge after routing. Therefore, we adopt different procedures for large and small basins to achieve maximum consistency in model-observation comparisons.

• P6/L149 "catchment area": which area estimates did you use? For all models the same? Per model based on catchment delineation? Please clarify to avoid that data was used inconsistently.

We only used the reported catchment area of each stream-gauge in this calculation. We will clarify about this technical aspect in the revision.

• P7/L177-L185: great you are pointing out the differences in methodology!

Thank you for your encouragement.

• P11/L228-L230: these lines read as if they should not be part of the methodology, rather of the results/discussion section. The fact that that there may be 'hot-spots' of future flood hazard should be discussed in more detail and thus deserves a more prominent location in the manuscript.

We appreciate the reviewer's suggestion. We will revisit this paragraph and consider highlighting the "hot-spots" aspect more in the revision.

• P11/L231 "each grid-cell": each grid-cell or only those paired with a GSIM-location? Does not become very clear from reading.

This analysis was conducted for each grid-cell across the globe, regardless there is stream-gauge or not. We will clarify about this to avoid confusion.

• P11/L235-L237: why was this done? What does it add?

This step was included to assess whether the locations that robustly projected with increasing/decreasing trends in flood hazard (i.e. the magnitude of MAX7 index increases/decreases significantly during the 2006-2099 period) has been observed adequately by the current streamflow observation system.

We will revisit this section in the revision to improve clarity.

• P11/L240-L247: You describe the observation and simulations, but you do not mention the reasons behind it. Why are certain areas experiencing increases and others decreases? Is it all hydrology or not? Can we say something about the driving factors behind it? Please add.

We acknowledge the reviewer's perspective about the importance of attributing changes in flood hazards to hydrological or climatic mechanisms factors. We noted, however, that the key objective of the present study is to assess model capacity rather than exploring the mechanisms driving changes in flood hazard. The reviewer also noted that the manuscript has been quite complex in its current state already. As a result, we propose to not include these discussions in the revision. Instead, we will make our objectives clearer, and will clarify that the paper does not focus on explaining the mechanisms driving changes in floods.

• P12/L255+L256: What is the implication of this finding?

We discussed about the implication of this finding at line 290, in which we suggested that averaging will reduce the magnitude of trends and thus ensemble average should not be used as a sole ground to infer change in floods. This is also a motivation for us to provide the range of trend characteristics across all ensemble members.

We will revisit our discussion to communicate this implication clearer.

P12/L261+L262: So, if it is not visible through Figure 2, how can it be an alternative explanation? This sounds contradictory to me – either it's possible based on your results or your results say it's not a thing. Please clarify.

The intention of this statement was to indicate that we would explore GHM's performance in more detail (i.e. through the next paragraphs/sections) because Figure 2 alone was insufficient to explain such feature.

We found this statement may be confusing and will revise it to improve clarity.

Figure 2: A bit bigger figure (maybe with subplots of USA, EU) would help seeing the differences between differences between historical trends.

We note that the Supplementary has included sub-region plots for this figure. This figure is also useful to highlight the "white spaces" over many regions, which was then linked to our call for more streamflow observation.

We will explore the options to improve graphical quality of this figure in the revision. Some possible options are to include the vector graphic or use another colour pallet.

• P14/L285-L287 and Table 3: what are possible reasons for the different model results? Model structure, processes simulated per model, spatial resolution, routing schemes applied or something else? Would be great if you could elaborate a bit on this.

From our perspective, this question is not straightforward to answer due to the number of participating models (six) and the many factors involved (e.g. the individual and collective effects of differences in model conceptualization, spatial resolution and routing scheme). These aspects (i.e. possible reasons for different trends simulated by different GHMs) in fact is s till under-represented in the literature. Even when model differences are documented extensively, it is still challenging to precisely attribute output discrepancies to a specific (or a set of) factor(s) without supports from another set of GHM simulations (e.g. checking the sensitivity of simulated trends corresponding to changes in a specific factor). Nevertheless, we will consider to elaborate about potential sources of differences in model outputs in the revision. We will also highlight this in the conclusion as a potential research pathway.

• P14/L290+L291: if it should not be used as "sole ground", what other measures would you (like to) use to infer changes in floods?

We will clarify in the revision that the range of all ensemble members should be used to illustrate the spread of simulated trends (e.g. the information showed in Table 4).

• P23/L463+L464: Does that mean your results are not usable to help inform flood management practices in less well-observed areas? What would be the implications? Please elaborate briefly on the consequences of your results.

The intention of this statement is to set the stage for our next analysis which shows the regions projected with increasing flood hazards are under-sampled, and ultimately leads to our call for more attention to improved streamflow observations. We will revise our discussion to improve the narrative of these ideas.

• P24/L475-L478: What would be ways forward to reduce the dependency on not evenly spatially distributed observation systems? Which opportunities do, for instance, remotely sensed data products bring? Please put your findings into context, here and/or the conclusions section.

This statement was used as a ground for our call (at the conclusion) for more FAIR streamflow observations to support hydrological research. We acknowledge that the narrative may need improvement and will revisit the paper, potentially including some of the reviewer's suggestions.

P26/L533-L537: It should be added that also the routing schemes of GHMs should improve, not only the runoff. Well timed runoff with right magnitude can still result if inaccurate streamflow if the routing scheme is too simplistic. Vice versa, higher-order routing schemes cannot perform at their best if input runoff is not accurate. Relevant literature: Hoch et al., 2019 (https://doi.org/10.5194/nhess-19-1723-2019) and Zhao et al., 2017 (https://doi.org/10.1088/1748-9326/aa7250).

We will extend our discussion to include the importance of the routing scheme on GHMs' performance and the need to improve this important feature in future GHM generations.

P27/L550-L559: I very much agree with this, well written!

We thank the reviewer for their encouraging comment.

References

Masaki, Y., Hanasaki, N., Biemans, H., Schmied, H. M., Tang, Q., Wada, Y., Gosling, S. N., Takahashi, K., and Hijioka, Y.: Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado, Environmental Research Letters, 12, 055002, 2017.

Zhao, F., Veldkamp, T. I., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., and Portmann, F. T.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environmental Research Letters, 12, 075003, 2017.