Dear authors,

First of all, I want to congratulate you for this extensive piece of work. It clearly shows that a lot of work was carried out to analyse this huge load of simulated and observed streamflow data.

While I think that the content fits within the scope of HESS and the work makes use of recent literature, models, and data, the manuscript needs improvement before publication is possible.

Please see below my general and subsequently my specific comments.

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?
- 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.
- 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).
- 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.
- 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.
- 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.

Specific comments

- P1/L29+L30: An agreement of 12-25 % can hardly be named “moderate agreement”, can it?
- P1/L31 “significant differences”: please specify what kind of differences you refer to.
- P2/L52 “specific regions”: specific regions such as? Please specify.
- P2/L53 “recent evidence”: what evidence? Please provide name, source, etc.
- P2/L52-L66: what about the role humans play in changing flood hazard? Please add this dimension to the paragraph.
- P3/L77 “factorial experiments”: what are you referring to with this term? Please explain or use more common terminology.
- P3/L87-L100: The description of the research objectives could profit from using bullet points
- P4/L105-L108: why are these models used? Why are some others available within ISIMIP not used? Please clarify.
- 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.
- 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?
- 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.
- P7/L177-L185: great you are pointing out the differences in methodology!
- 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.
- P11/L231 “each grid-cell”: each grid-cell or only those paired with a GSIM-location? Does not become very clear from reading.
- P11/L235-L237: why was this done? What does it add?
- 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.
- P12/L255+L256: What is the implication of this finding?
- 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.
- Figure 2: A bit bigger figure (maybe with subplots of USA, EU) would help seeing the differences between differences between historical trends.
- 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.
- 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?

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