Reviewer #2 Comments

Review of "Assessing the Hydrological Impact Sensitivity to Climate Model Weighting Strategies " by Asenjan et al.

- **General comments**

The paper is generally well written and presented but suffers from the use of terms and assumptions that ruin the validity of the work that was carried out. Furthermore, many choices made lack justification and a discussion on the alternatives available.

While I acknowledge the value of using a GCM as “pseudo-reality” I find the alleged aim “to seek accuracy and reliability in future hydrological runs” an outrageous claim. Even making a clear disclaimer at the beginning of the paper (which is not there), these terms are simply not applicable in this context. I think the authors should have approached this work under an uncertainty reduction effort type of exercise as opposed to an accuracy effort.

Moreover, there are important concepts that the authors have neglected: 1. model weighting schemes are subjective and not necessarily overcome the 1 model 1 vote (model democracy) – a discussion on this would have provided a more solid basis to their work 2. Assuming that a model that performs well in the control period should be deemed more credible in the future period than other models that did not is an assumption that is debatable and should be at least discussed. 3. If a few runs are “outliers” it doesn’t necessarily mean that they will get it wrong, the other majority could be wrong too, that is why model weighting is a tricky job and cannot live without a comprehensive assessment of uncertainty associated.

For these reasons I think this paper is unfit for publication.

I have noted specific comments below.

Thank you for your thorough review and feedback. We appreciate your critical assessment and the time you have taken to provide detailed comments. We respectfully disagree with the conclusion that the paper is unfit for publication. Instead, we see the issues raised as opportunities for significant improvement and clarification.

We are committed to addressing all the specific comments and concerns raised. Specifically, we will:

1. Revise the use of terms like "accuracy" and "reliability" to focus on appropriate terminology and uncertainty reduction.

2. We will also add a discussion on the subjectivity of model weighting, discuss that good model performance may not necessarily translate to good performance in the future, the importance

of a comprehensive uncertainty assessment, and include that outliers may still add valuable information and may not be completely wrong.

3. Add discussions and citations where necessary to justify our methodological choices and acknowledge the limitations and alternatives.
4. Clarify the objectives and contributions of the study, ensuring they are realistic and accurately reflect the scope of our work.

We believe that with these revisions, the paper will provide valuable insights and contribute meaningfully to the ongoing discussion in hydrological and climate modeling research. We hope you will consider our revisions and reassess the paper's suitability for publication.

You will find a point by point answer to your comments below

- **Specific comments** (section addressing individual scientific questions/issues)

16 - "Various weighting schemes address these concerns, but their effectiveness in impact studies, which integrate GCM outputs with separate impact models, remains unclear." Their effectiveness will remain unclear until the future unfolds and we find out what actually happened.

This is of course the main reason for our methodological choice of using the 'virtual world' approach, as it provides a future 'truth' against which to test weighing schemes. We will revise the sentence to better reflect this statement. Do note that the revised version will feature an enhanced section on the 'virtual world' as suggested by reviewer 1 comments, to better account for the benefits and limitations of this choice.

35-36 – Accuracy and Reliability are concepts that pertain to weather forecasting and sub-seasonal to seasonal prediction, I find it misleading to use them in the context of climate/hydrological projections.

Thank you for raising this point. It is valid and we will modify the manuscript to avoid confusion with forecasting.

48 – "This variability is widely recognized as a primary source of uncertainty". This variability is only fair and must exist if we want to describe and account for the plausible future of the variable of interest.

We fully agree with this statement. We will emphasize this better in the revised version.

57 AND L. 60-70 – I think it is more complex than that. "While model democracy has been successful in replicating the mean state of the observed historical climate (Reichler & Kim, 2008), its applicability and reliability in future impact assessments remain uncertain." Broadly speaking model democracy makes no judgment on the different models of the ensemble, while weighting or excluding runs implies some sort of judgment that is not easy to justify and often ends being, to some degree, subjective. In particular, not everyone agrees with the fact that if a model replicates decently a given variable in the control period, it will continue to do so in the future. Therefore: "suggesting model democracy might not be the

best choice in regions where some models are more reliable" – this "more reliable" is a crucial concept in this issue of model democracy – vs. – model selection/weighting because measuring reliability of future quantities simply isn't possible until the future unfolds.

One may find indications based on assumptions, but I think the ease with which some concepts are reported and treated is misleading (accuracy and reliability).

We fully agree with this statement, and the methodology we used in this paper allows us to test these assumptions. In retrospect, we agree that the introduction (and discussion) could have delved into this topic in more detail, and we will do so in the revised version.

We acknowledge that weighting GCMs involves subjective decisions regarding the weighting criteria and methods. At the same time, model democracy is also a subjective choice, which is the focus of this paper.

By "more reliable" (a term we will not use in the revised version), we meant models that perform better during the observation period in reproducing critical climatic processes in the studied region. 'Reliability' may not simply relate to past performance against key GCM outputs such as precipitation and temperature but may also be based on model physics. For example, it may be reasonable to weigh climate models with more complex ocean physics more heavily for a climate change study in Japan and Hawaii, or to exclude a GCM (weight of 0) that does not physically represent the Great Lakes for a study over Western New York State. These choices always remain partly subjective, but this is what the 'model-as-truth' experiments are designed to test. We will clarify this point in the revised version by expanding the introduction and discussion.

73-74 – "demonstrating more accurate projections compared to simple averaging". Same as above: how can you "demonstrate" accuracy in the future?

This is the whole point of using the 'model-as-truth' experiment. Once you take a GCM as 'truth,' you can test weighting hypotheses since you have future data. As mentioned above, based on comments by Reviewer 1, we will better explain the strengths and limitations of this approach.

91 – I would add Giuntoli et al. 2021 on the effect of weighting via streamflow data.

We will, thanks !

97 – On Pseudo observed – model as truth, imperfect model test. Please explain to the reader that the limited number of simulations involved in earlier studies did not allow to separate the inference of internal variability from structural differences among the models. With large-ensemble initial condition simulations there is the advantage of having many simulations that can be used as pseudo-observations (Deser et al. 2020).

We definitely will, along the expanded discussion on the model-as-truth approach.

115 – “significance” – please avoid terms that have to do with statistical testing.

We feel silly missing this as this is a point we constantly raise when acting as reviewers. We will correct this accordingly to avoid confusion.

119 – “the objective is to understand the complex interactions between weighting schemes and their effects on […]” – I don’t think this study allows to understand the complex interactions between weighting schemes, it simply conducts a sensitivity with a bunch of them. I suggest more realistic words for the objective.

Agreed. We will revise the text accordingly.

123-127 – Again, I struggle with the terms “accuracy” and “reliability” in the context of hydrological projections. I would either write a clear disclaimer in the introduction, or would consider using other terms.

Answered above. Definitely a misuse of these terms.

131 – The data set includes ~14k catchments from which ~3k were chosen. The selection criteria are debatable. See Giuntoli et al. 2015, where catchments are selected to be of comparable size to the gridcell resolution of the global models. Therefore a minimum area of 500 km2 is fair, to avoid flashy catchments, but there should be an upper limit too.

We selected the catchments based on the hydrological model performance, ensuring that the HMETS model achieved a minimum KGE of 0.5 after calibration. Additionally, we used a minimum threshold of 500 km² for the catchment size to avoid flashy catchments. While we did not impose a maximum area limit, it is worth noting that 97.5% of the selected catchments have areas less than 11,000 km², with the largest catchment area being 650,000 km², which is in a similar range to Giuntoli et al. (2015). The regional consistency in the observed results does not suggest any important effect of catchment size. We will however discuss these issues in the revised manuscript.

L.138 – perform comparably to observation data in hydrological modelling is very vague. It really depends on the variable and the location. This point should be further justified and described. I would also add citation on the assessment of ERA5 precipitation by Lavers et al. 2022.

ERA5 performance does indeed vary by variable and location. ERA5 precipitation shows reliable performance in extratropical regions, which aligns with our study area (Lavers et al., 2022). Additionally, Tarek et al. (2020) demonstrated that hydrological modeling based on ERA5 performs as good as using observational data over most of the USA; however, observations provide better performance in the eastern US. We will update the manuscript to reflect these nuances and include the relevant citations.

145 – Figure 1, what is the purpose of indicating mean annual temperature?

The purpose of including mean annual temperature in Figure 1 was to illustrate the spatial distribution of selected catchments and their associated temperature characteristics. This helps to show how catchments vary in terms of their hydrological context, highlighting differences in temperature across the study area. Just showing basin contours appeared to us as a wasted opportunity. In the revised version we propose to revise Figure 1 by changing its color coding to reflect the performance of the hydrological model. This change will enhance the figure's informativeness by directly linking model performance with catchment characteristics. We believe this will provide a better hydrological assessment.

170 – In addition to the reference for the method (Cannon 2018), there should be an explanation for using bias correction based on pseudo-reality GCM data with references and limitations.

Thank you for your comment. We will address this in the revised version. We will emphasize that bias correction in this study was conducted using pseudo-reality GCM data. We will discuss the main limitations of bias correction as well as its necessity in certain cases. We will, of course, add supporting references. Some of these elements are already present in the text, but we will add significant additional information.

175 – What is the spatial and temporal resolution of the HMETS model? There should be a description of the resolution at which GCMs are input into HMETS and the resolution that HMETS outputs.

We will clarify that the HMETS model is a conceptual lumped hydrological model that operates on a daily time scale, with both inputs and outputs at this temporal resolution. We will also add a brief discussion of the spatial scales of GCM grids and catchments size distribution.

183 – Six weighting schemes are adopted. Why? What led you to this choice? Review of other studies employing weighting schemes? E.g. in the hydrological community it has been shown how far off hydrological models fed by GCMs are from observations, in particular with regards to extremes (e.g. floods). Some have attempted weighting the models (or selecting them) on the basis of their ability to reproduce the right timing of hydrological events as opposed to the quantity (m3/s) e.g. Giuntoli et al. 2021.

Thank you for your comment. We selected the six weighting schemes based on recent literature to encompass a range of criteria, including GCM performance, model independence, and random and equal weighting for comparison. We focused on weighting schemes based on either precipitation and temperature, as they are the most common, and the one used by the climate modeling community,

We will revise the manuscript to provide a better justification for the selection of these weighting schemes, as well as providing a brief review of other possible weighing schemes (including ones based on hydrology) that could have been used. Ultimately, we believe that our conclusions would be the same if we had used others schemes, and this will be tackled in the revised discussion.

188 – Monthly observed and simulated series. Is there a downgrading of the data from daily to monthly temporal scale? Also, "observed" in this case is the pseudo-reality series? If so, I find using "observed" confusing for the reader.

Yes, we do aggregate the daily data to a monthly temporal scale to apply the equations used in our analysis.

We apologize for the confusion. In this section, we used "pseudo-reality" data. "Observations" are solely used for calibrating the hydrological model. We will rephrase "observed" to "pseudo-reality" in the manuscript to avoid any ambiguity and ensure that all terminology and data descriptions are precise.

195 – assessing how a GCM aligns with the multi model mean in future projections. I think the assumption made by this metric is wrong: a model that does not align with the majority of the models is deemed less credible. What if that model is the only one that gets it right?

This is a valid point and we agree with it. However, the climate modeling community has traditionally been suspicious (to a certain extent) of 'outliers' and this metric represents this. This is obviously a subjective choice which goes back to the points discussed earlier. The modifications suggested earlier will encompass our choice of metrics. It is noteworthy that the paper on reliability averaging weighing (Girogi and Mearns, 2002) has been cited nearly 1000 times so it is a very common method in the climate modeling community.

Again, observations are mentioned for the REA metric, are these observations or pseudo-reality model runs?

We apologize again for the confusion. The term "observations" in this section was intended as a general descriptor rather than an accurate reflection of our methodology. We used "pseudo-reality" data for the analysis. As discussed above, we will rephrase this in the manuscript to avoid any ambiguity and ensure clarity.

211 – Skill metric. This metric implies that models that reproduce observed data the best are to be trusted more in their projections. This assumption should be clarified and discussed because it is not necessarily true that models that do well in the past will continue to do so in the future. To some extent this characteristic increases their credibility, but does not ensure improvements in future performances.

Agreed. This has been discussed above, and the revised manuscript will have enhanced introduction and discussion to discuss these aspects.

265 – On the need to bias correct GCM output. Please justify this choice further, mentioning drawbacks for bias correcting, e.g.: the introduction of an additional source of uncertainty; the reduction of inter-GCM variability. I would cite Ehret et al. 2012.

This also goes back to an earlier comment. We will do so in the revised manuscript. We will also cite relevant literature, including Ehret et al. (2012) - thanks for the suggestion.

272 – It seems like there is one and only hydrological model that routes streamflow using precipitation and temperature. As hinted above, descriptions of other options (models) available is needed.

Indeed, there are a plethora of options. We will expand the section on hydrological modeling and enhance the discussion part of the limitations of only using 2 lumped conceptual hydrological models.

281-282 – Again, the use of the term accuracy is far-fetched. “accurately” capturing the key underlying hydrological processes is wishful thinking, unattainable with the plethora of GCMs run at coarse resolution, then bias-corrected, then fed to a hydrological model that uses P and T as input. The cascade of uncertainty is such that the accuracy you look for is simply not there. You can put effort on trying to minimize uncertainty, but cannot write about accuracy here.

Agreed and discussed above.

283-285 – I don’t understand this statement: “as long as the processes are reasonably represented (are you assessing this?) the model performance is not of critical concern”. Which model? And why it is not a concern?

Thank you for your comment. We apologize for the confusion; we meant ‘hydrological’ model. We will ensure that we always specify ‘model’ throughout the text. We calibrate the hydrological model based on the ERA5 data and select only catchments with a Kling-Gupta Efficiency (KGE) above 0.5, indicating satisfactory hydrological model performance. Modeled streamflow biases will be present to some extent, but we can assume that these biases will persist in the future period. Since the paper focuses on climate model weighting, we don’t believe that streamflow modeling biases have a significant impact. To support this, as mentioned in the text, we used a second hydrological model, which did not alter any of the conclusions. We will expand the discussion on this in the revised version.

291-292 – Why are the two methods supposed to yield similar results?

Since after applying bias correction, the streamflow characteristics of the 21 GCMs (GCM_k) should closely resemble those of the pseudo-reality GCM_i, different weights do not necessarily result in a weighted average that is significantly different from the pseudo-reality. We will expand on this in the revised version.

References:

Giuntoli, I., et al. (2021). Going beyond the ensemble mean: Assessment of future floods from global multi-models. Water Resources Research, 57, e2020WR027897. https://doi.org/10.1029/2020WR027897

Deser, C., et al., 2020: Insights from Earth system model initial-condition large ensembles and future prospects. Nat. Climate Change, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2

Giuntoli, I., et al. (2015), Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States, J. Geophys. Res. Atmos., 120, 9138–9159, doi:10.1002/2015JD023401

Lavers, D.A., et al. (2022) An evaluation of ERA5 precipitation for climate monitoring. Quarterly Journal of the Royal Meteorological Society, 148(748) 3124–3137. Available from: https://doi.org/10.1002/qj.4351

Ehret, U., et al. (2012), J.: HESS Opinions “Should we apply bias correction to global and regional climate model data?”, Hydrol. Earth Syst. Sci., 16, 3391–3404, doi:10.5194/hess-16-3391-2012

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., & Schmidt, G. A. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. Earth System Dynamics, 10(1), 91–105. https://doi.org/10.5194/esd-10-91-2019

Chen, Z., Zhou, T., Chen, X., Zhang, W., Zhang, L., Wu, M., and Zou, L.: Observationally constrained projection of Afro-Asian monsoon precipitation, Nat. Commun., 13, 2552, https://doi.org/10.1038/s41467-022-30106-z, 2022.

Giorgi, F., & Mearns, L. O. (2002). Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the “Reliability Ensemble Averaging” (REA) Method. Journal of Climate, 15(10), 1141–1158.

Maraun, D., 2012. Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums: NONSTATIONARITIES OF RCM BIASES (n/a-n/a). Geophys. Res. Lett. 39. https://doi.org/10.1029/2012GL051210.

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., & Menary, M. B. (2023). Performance-based sub-selection of CMIP6 models for impact assessments in Europe. Earth System Dynamics, 14(2), 457–483. https://doi.org/10.5194/esd-14-457-2023