Reviewer #1 Comments

Summary:

The manuscript titled "Assessing the Hydrological Impact Sensitivity to Climate Model Weighting Strategies" evaluates the impact of different weighting strategies for GCM outputs on streamflow projections. The study uses the concept of pseudo-reality to address the challenge of lacking future observations.

We thank the reviewer for taking the time to review our paper, recognizing its strengths, and providing valuable feedback for improvement. You will find below a point by point answer to your comments.

Strengths:

1) The manuscript provides a comprehensive analysis of various weighting strategies, which could be helpful for further studies.

2) The use of pseudo-reality to overcome the lack of future observations can be useful and, if well-explained, could provide insights into model weighting strategies.

We are pleased that you found our comprehensive analysis of various weighting strategies helpful for further studies and appreciate your recognition of the potential usefulness of pseudo-reality to overcome the lack of future observations. In the revised manuscript (also taking into account comments from Reviewer 2), we will elaborate on this concept to enhance clarity and understanding.

Weaknesses:

1) However, the actual scientific contribution of the paper is difficult to assess as the research gap is not clearly defined. For example, a specific aim of the study is to investigate how the selection of assessment criteria influences the results of climate model weighting in hydrological assessments (line 125), while it seems that this question has been well discussed in the literature (lines 90-95). Although it is further mentioned that it is challenged by the lack of future observations and can potentially be addressed with pseudo-reality, the research questions are still not clear. What is the state of the art and what are the implications of the study regarding the use of pseudo-reality for weighting schemes?

Thank you for this comment. We understand that considering the current introduction, it can be difficult to recognize the main research question of the study. The main goal of our study is to fill a critical gap in understanding how different climate model weighting strategies impact hydrological assessments, particularly in the context of future climate projections where observational data is unavailable.

Furthermore, our study addresses several key aspects that remain underexplored. Firstly, considering that there is still no consensus on the most effective weighting method, our study, using the pseudo-reality approach, provides a thorough evaluation of different weighting schemes, highlighting their

relative effectiveness and potential implications for future climate impact studies. This contributes to the ongoing debate on model weighting effectiveness.

Secondly, the use of pseudo-reality (model-as-truth) experiments allows us to test the robustness of various weighting schemes without the limitations posed by the absence of future observational data. This approach provides a novel way to benchmark and compare the performance of different weighting methods in a manner that is not possible with real-world data alone. By identifying weighting schemes that perform well in pseudo-reality tests, we can improve future hydrological projections, which is crucial for effective climate adaptation and mitigation strategies. By treating pseudo-reality as a reference, we can rigorously test the effectiveness of these schemes in projecting future hydrological impacts, thereby providing a more robust validation framework.

Finally, our study employs multiple iterations of the pseudo-reality method, considering various climate variables and geographic regions. This comprehensive approach enables a more nuanced understanding of the sensitivity of different weighting schemes, thereby offering deeper insights into their implications for future climate impact studies.

We will modify the document to clearly emphasize these points to ensure the research gap is clearly defined.

2) The rationale for using pseudo-reality to address the lack of future observations is not well-explained. More detailed justification and explanation of the hypothesis behind pseudo-reality are needed, and citations alone are not sufficient to justify the effectiveness of the practice in the context of this study. Also, are there alternatives to pseudo-reality and how can it be ensured that the study's conclusion about the effectiveness of weighting is not just a product of this particular technique? Some further discussion is needed.

In retrospect, we acknowledge that our paper could have covered the basics of the pseudo-reality approach more comprehensively. The use of pseudo-reality, or model-as-truth experiments, is based on the need to overcome the lack of future observational data. It provides a controlled environment to evaluate and benchmark different weighting strategies without the confounding factors present in real-world data. More specifically, it provides a known reference against which different weighting schemes can be benchmarked. This helps in assessing the performance and robustness of these schemes in a systematic and controlled manner.

To further elaborate, the model-as-truth approach involves removing one ensemble member and treating it as if it were observations. The remaining ensemble is then compared against this “truth” member since the “truth” member’s projections are known. This process is repeated with each ensemble member playing the “truth” role. The ability of the weighting to offer improvement over the original default ensemble is then assessed across all model-as-truth cases to gain an understanding of the efficacy of the particular approach being tested.

The underlying hypothesis of using pseudo-reality is that if a weighting scheme can accurately replicate the pseudo-reality scenario, it is likely to be effective in projecting future climate impacts. By using multiple pseudo-reality scenarios, we simulate the range of uncertainties inherent in climate projections, helping to identify weighting schemes that are consistently reliable across different scenarios.

As with any method, there are limitations to the approach. Being in a 'climate model' world, the model-as-truth approach may oversimplify complex real-world processes, potentially missing important factors that influence climate impacts. In addition, without real-world future observations, results remain theoretical despite the advantages described above.

While pseudo-reality is a powerful tool, it is not the only method for evaluating climate model weighting. An alternative approach involves using a traditional calibration-validation framework with historical data partitioned into calibration and validation sets. This allows for testing the applicability of weighting schemes out-of-sample using past climate data. Ensemble averaging has also been used as a future target. Using expert judgment or advanced techniques such as Bayesian averaging has also been suggested and used in the literature. However, all of these methods have limitations, including an important common one which is the absence of future climate data without which it is extremely difficult to truly benchmark weighting methods since climate model sensitivity is overlooked. For example, a climate model with excellent performance on historical data could have an equilibrium sensitivity way too high (or too low) and may not deserve the high weight obtained over the historical period.

We will add the above points (along with additional relevant literature citations) to the manuscript. In particular, we will have a complete sub-section to discuss the pseudo-reality approach.

3) The manuscript's presentation needs improvement. For example, some statements lack references (e.g. lines 100-105), and some paragraphs need to be more coherent (e.g. the last two paragraphs in the introduction are repetitive, where many aims/goals/objectives are presented). Also, some tables and figures are confusing. For example, the meaning of the ID number in Table 1 is unclear, and it appears just the one used in Figure 2. What is the order of the climate models in Figure 2? Is it just random? If so, the figure is not informative and I would suggest that it be ordered by ECS or some similar metric to make it meaningful.

We will improve the presentation of the manuscript by adding references to statements that lacked citations (e.g., lines 100-105).

We will also revise the introduction to improve coherence and remove repetitive paragraphs. The aims, goals, and objectives will be clearly presented in a single, concise paragraph.

Regarding the model IDs, they are alphabetically ordered by the modeling center but listed by ECS, hence the confusion. We will re-number them using ECS to avoid this issue.

Specific questions:

- Line 156: what specific statistical metric is being compared?

We compare projected temperature change (the difference between the future and reference periods) and precipitation change ratio (calculated as (future P - reference P) / reference P) over all 3,107 catchments for all 22 GCMs.

We will explain this more clearly in the revised manuscript..

- Line 159: how is the choice of weighting model related to climate sensitivity here, i.e. why can the finding of different ECSs highlight the potential importance of the choice of weighting model? Please clarify.

It has been suggested that GCMs with higher Equilibrium Climate Sensitivity (ECS) may present less realistic or less probable future scenarios *(Hausfather et al., 2022).* Therefore, these models should be either removed *(Rahimpour Asenjan et al., 2023) or down-weighted (Massoud et al., 2023)* when studying ensemble means. Consequently, ECS becomes a critical factor in the weighting of models. This relationship highlights the importance of carefully choosing the weighting model to ensure that the projections are both realistic and scientifically robust.

We will add this point to the manuscript.

- Lines 169-170: Please justify the operation. Why is this not done using the corresponding GCM data?

Bias correction was performed exclusively using pseudo-reality GCM data. This approach means that the pseudo-reality is considered the truth, and GCM data is corrected based on this truth. In hydrological impact assessment studies, GCM data is typically bias corrected using available observations. Here, the pseudo-reality serves as the hypothetical reality, so bias correction is done based on this considered as truth models (similar to Maraun, 2012), ensuring that the corrected GCM data aligns more closely with the considered reality.

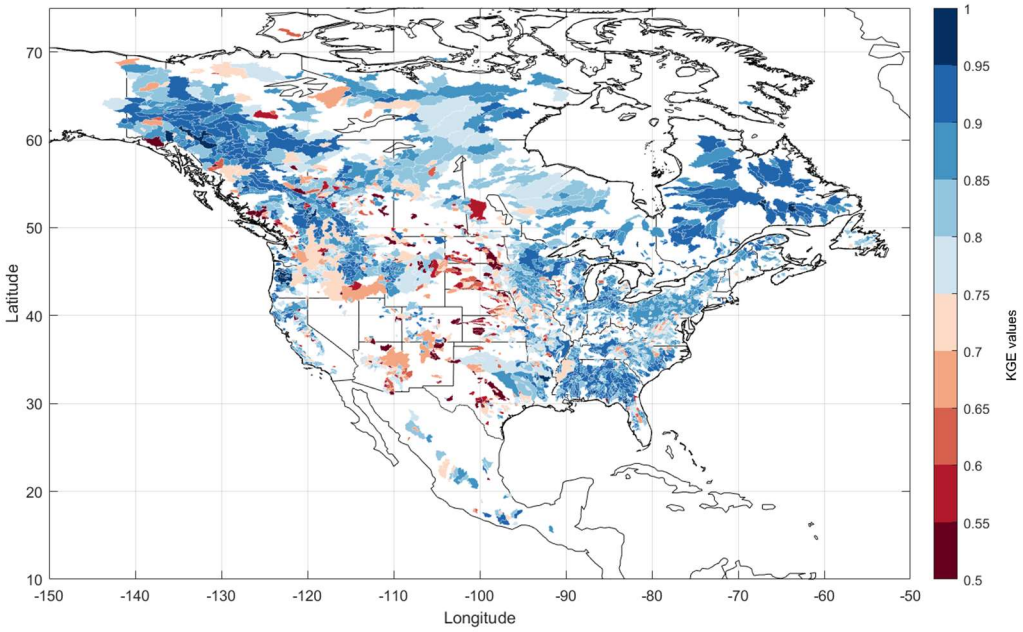
We will revise the text to improve clarity and provide a clearer justification for this methodology.

- For the HMETS model, no performance seems to be reported in the paper.

We apologize for the oversight. We will include the following sentence at the end of section 2.2 to report the performance of the hydrologic model:

"The selected watersheds all have a Kling-Gupta Efficiency (KGE) value of over 0.5, with a median and maximum values of 0.83 and 0.97 respectively. These values indicate a satisfactory model performance over all selected catchments."

Additionally, considering Reviewer 2's comments, we will change the color coding of Figure 1 to reflect the performance of the hydrological model, making it more informative. The revised figure is provided below.



"Figure 1. Map of the 3,107 catchments used in this study. The color code represents the hydrological model Kling–Gupta efficiency (KGE) calibration score over the reference period. In the case of nested catchments, the smaller ones were plotted on top of larger catchments."

- Line 271: Yes, I agree, but how are the GCM meteorological forcing matched to the lumped hydrological model in the study, where the minimum basin size is 500 km2.

The GCM meteorological forcing data is matched to the lumped hydrological model by averaging the values of all grid points within the watershed outline. If grid points are found within the watershed outline, their values are averaged to obtain a representative value for the watershed. If no grid points are found within the watershed, the closest grid point is used. We will add this to the revised manuscript.

- Line 276: The statement is still not referenced and the rationale for the practice is not explained.

Thank you for bringing this up, since this sentence touches on the pseudo reality approach, we will add relative references and the explanation mentioned above (Main point 2), to enhance clarity of the manuscript

- Line 281: I am more confused here. Does this mean that pseudo-reality is assumed to be more accurate than ERA5 data? I am curious how the streamflow prediction performs with the pseudo-reality historical scenario? If it needs bias correction itself, how can it be used as a reference for other GCMs? Please clarify here.

The pseudo-reality is not assumed to be more accurate than ERA5 data. Instead, pseudo-reality serves as a hypothetical construct to test model weighting strategies. There is no comparison regarding the accuracy of ERA5 and pseudo-reality; ERA5 data is only used for the calibration of the hydrologic model to ensure that the hydrological model performs adequately for the representation of streamflow given hydrometeorological data.

The streamflow prediction using pseudo-reality depends on the time series generated from the pseudo-reality data. This prediction is not meant to be directly comparable to observation data, as GCMs aim at representing the climate and cannot represent the exact daily observed time series.

The pseudo-reality serves as a reference for other GCMs because it is treated as the hypothetical "true" scenario. The bias correction of GCM data using pseudo-reality ensures that the corrected GCM data aligns more closely with this considered reality, allowing for a consistent comparison across different weighting strategies.

As mentioned in Section 2.4, we consider each GCM as a potential pseudo-reality, not just one. This approach allows us to test the effectiveness of different weighting strategies in a controlled manner. If an averaging method performs better in the pseudo-reality scenario, it suggests that it could perform better with real data as well.

We will clarify this in the manuscript to ensure a better understanding of the methodology and its purpose.

- Lines 310-312: why red coloring suggests better performance than equal weighting. The metric here is bias, so my understanding is that it depends on the estimation errors of the equal method. If the equal method already has a negative bias (i.e. red in Figure 4a), doesn't the negative difference in Figure 4b-f (i.e. red) imply an even worse case? I hope I have just misunderstood something.

The metric used in Figure 4b-f is the difference in the absolute values of bias. This means that if the equal weighting method already has a negative bias, its absolute value would be positive. If the difference between the absolute values is negative, it indicates that the other method is performing better by having a lower absolute bias value.

The text will be improved for clarity.

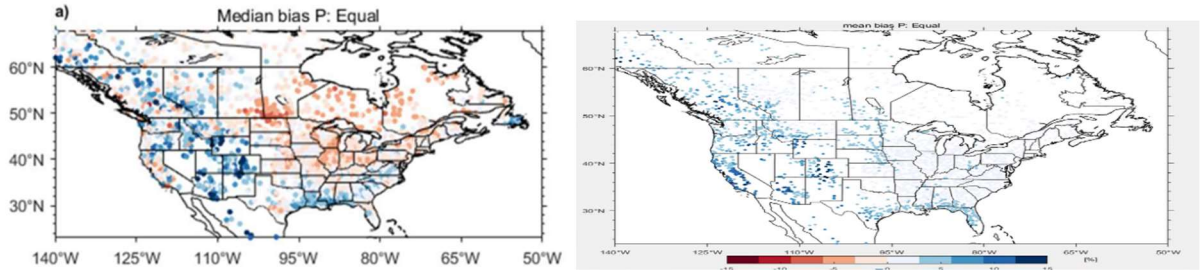
- Lines 326-329: Is it possible to explain why the spatial distribution pattern is as it is, i.e. why western and south-eastern catchments have a negative bias? Understanding the underlying reasons will be helpful in knowing what potential geographical factors are being referred to here.

Figure 4 (and others) presents the median bias. The median value is taken from the distribution of bias resulting from the 22 model-as-truth experiments (taking each of the 22 GCMs in turn as the 'truth'). Negative median biases indicate a right-skewed distribution, while positive biases indicate the opposite.

The two figures below show the median (left-hand side, same as Figure 4a of the original manuscript) and the mean (right-hand side). Both figures use the same color scale. We can see that the biases are much smaller when looking at the mean. This suggests that some climate models are strongly negatively biased in the west, whereas the opposite, to a lesser extent, is true on the east coast. The underlying reasons for this are not clear.

However, it is notable that regions with a positively biased median (in red) all have a continental climate (Dfa, Dfb, Dfc, and Cfa of the Köppen climate classification), whereas regions with negatively biased medians have either maritime or mountainous climates. A similar difference in behavior was observed in ERA5 precipitation (e.g., Tarek et al., 2020) with negative precipitation biases in the same zones. This perhaps suggests some shortcomings in the model physics for certain types of climate.

To delve into this issue in more detail, we would need to examine the distribution of all 22 values of biases to determine if specific climate models are particularly skewing the results of the median bias. However, we believe this is slightly off-topic and would make the paper less streamlined. We will briefly discuss these issues but do not intend to go into further detail in the revised manuscript.



- Section 4.1: This is more of a repetition of the introduction. Might consider merging it with the introduction.

That's true, in the revised manuscript we will merge the redundant parts of Section 4.1 with the relevant parts with the introduction to streamline the manuscript and avoid redundancy.

We have also noticed that the results were mentioned as Section 4 instead of Section 3, and we will correct this.

- Lines 487-489: As an important conclusion, which is also highlighted in the abstract, it would be better here to articulate how it is supported by the results of this study.

We believe that the conclusion in lines 487-489 is well-supported by the results discussed in the previous sections of the study and we will make sure to better articulate the reasons behind this in the revised version.

Overall, while the manuscript provides a comprehensive analysis, it needs more clarity in its hypotheses, a clearer presentation and a better defined research gap to improve the overall quality of the manuscript. More scientific discussion (e.g. the robustness and applicability of the conclusion) is needed to enhance its significance and scientific contribution.

Thank you once again for your insightful feedback. We will revise the manuscript to improve clarity in hypotheses, presentation, and the definition of the research gap. We will also add more scientific discussion on the robustness and applicability of our conclusions to enhance the manuscript's significance and scientific contribution.

We believe the revised manuscript will address your concerns and improve the overall quality of the manuscript.

References

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M. (2022). Climate simulations: Recognize the 'hot model' problem. *Nature*, *605*(7908), 26–29. https://doi.org/10.1038/d41586-022-01192-2

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.

Massoud, E. C., Lee, H. K., Terando, A., & Wehner, M. (2023). Bayesian weighting of climate models based on climate sensitivity. *Communications Earth & Environment*, *4*(1), 1–8. https://doi.org/10.1038/s43247-023-01009-8

Rahimpour Asenjan, M., Brissette, F., Martel, J.-L., & Arsenault, R. (2023). Understanding the influence of "hot" models in climate impact studies: A hydrological perspective. *Hydrology and Earth System Sciences*, *27*(23), 4355–4367. https://doi.org/10.5194/hess-27-4355-2023

Tarek, M., Brissette, F. P., & Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrology and Earth System Sciences, 24(5), 2527-2544.