Review for „Emerging global freshwater challenges unveiled through observation-constrained projections” by Fei Huo et al.

General comments:

The authors use the EC (emergent constraint) methodology to observationally constrain TWS changes in ISIMIP3b and ISIMIP2b model output. The results show a substantial reduction in TWS when applying the constraints under different emission scenarios by the end of the century as compared to the raw model output, indicating that unconstrained model results might underestimate future water scarcity. While the overall findings seem reasonable to me, and pointing out shortcomings of TWS model projections is a relevant topic, I have some major concerns about this study:

Response: We appreciate your supportive assessment of our paper.

1. The EC method which is the backbone of the study does not get clear to me from the description in the manuscript. The text is lacking a clear introduction of the general idea behind the EC framework. Furthermore, advantages compared to other methods are not discussed.

Response: Discrepancies exist in future TWS projections, due to various factors, including uncertainties in climate forcing, the absence of key components such as surface water storage, groundwater storage, and human interventions in most land surface models (LSMs), as well as limited storage capacities within both LSMs and global hydrological models (GHMs). The EC technique, a relative novel method, is an efficient way to constrain uncertainties in future projections. Specifically, the EC technique aims to reduce the often uncomfortably large spread in future projections within multi-model ensembles, thereby providing more tightly constrained estimates of variables of interest (Hall, A., Cox, P., Huntingford, C. et al. Progressing emergent constraints on future climate change. Nat. Clim. Chang. 9, 269–278 (2019). https://doi.org/10.1038/s41558-019-0436-6). Such improved and physically informed constraints are crucial for climate mitigation and adaptation policymaking. We will modify the introduction based on your comments.

2. The study uses ISIMIP3b data, and ISIMIP2b data as “validation” data set. However, it is not explained in which way ISIMIP2b can be used for validation. In my opinion, the EC method is applied to both data sets, and no real validation has been carried out. In combination with the unclear description of the EC approach this leaves me with doubts about the validity and robustness of the results.

Response: ISIMIP3b and ISIMIP2b data were no validation datasets but necessary parts for the EC method, used to identify statistically significant relationships between annual global mean changes in TWS (y) and historical annual global mean TWSA climatologies (x), across a variety of global models. Such relationship (regression between y and x) along with the GRACE observations were then used to calibrate future TWS changes to reduce the spread in future projections within ISIMIP3b and ISIMIP2b models. Specifically, by replacing x with the GRACE observations in the regression, we obtained the calibrated future changes in TWS.

3. The investigation of underlying physical processes (Section 3.3) is not convincing to me. It must be extended and discussed in more detail.

Response: We will quantify regional biases and apply partial correlation analysis among different variables to get insights to understand mechanisms of TWS changes.

4. The authors do not discuss limitations of the EC method but see the challenges only in structural dependencies among climate models. Furthermore, uncertainties and limitations in the observational data and its implications on the results are not included in the discussion.

Response: The applicability of the EC technique is inherently limited by the knowledge space represented by the ensemble of climate models. If key physical processes are oversimplified or absent in the models, the EC method cannot identify or constrain those processes. Thus, the inter-model spread captured by the EC relationship may be unrealistic or unjustified due to the lack of a sound physical basis. We will improve the conclusion so that it could be better with the limitations and broader implications, including brief acknowledgment of potential limitations, such as short observational baselines, GRACE uncertainty, and non-stationary relationships under extreme forcing scenarios.

Overall, the manuscript would benefit from an independent validation of the method, a more elaborated explanation of the results, as well as a more critical discussion of the findings. Therefore, I recommend a major revision.

Response: We will enhance the manuscript by improving the methodology, the presentation of results, and the discussion.

Specific comments:

Section 2.1: It is not only GRACE but also GRACE-FO data being used. Please add GRACE-FO in line 49. You claim mascons being more reliable than spherical harmonics in the second sentence, but in the third sentence you mention that you also incorporate SH solutions. That is a bit confusing. I would suggest to base the analysis on the usage of all (mascon and SH solutions), and not to split it into mascon and mascon+SH, i.e., replace Fig. 1 by Fig. S7. This would be easier to read and follow.

Response: We will modify the text to improve its clarity. Since the inclusion of SH solutions has little impact on the outcomes, we have decided to remove the related discussion from the main text while keeping Fig. S7 for readers interested in the results that include SH solutions.

Line 70: "validated" What is the reasoning that ISIMIP2b can be used as a validation data set?

Response: The reviewer may have been misled by the use of this term "validated". As we mentioned in the response to general comment #2, ISIMIP3b and ISIMIP2b data were no validation datasets but necessary parts for the EC method, used to identify statistically significant relationships between annual global mean changes in TWS (y) and historical annual global mean TWSA climatologies (x), across a variety of global models. Therefore, we will improve the clarity in the revised version.

Line 73: Why regridding to 1x1° and not keep the 0.5° resolution?

Response: Some SH solutions such as GFZ RL06 and COST-G RL01 provide data only at 1-degree resolution. Thus, we regridded all datasets onto a common grid.

Line 75 / line 110: "GRACE's baseline period" sounds as if there would be a commonly defined period to which results always refer to, also in other studies. But I think this baseline was chosen by the authors specifically for this study? Please reframe.

Response: We will update the relevant content.

Line 78 – 85: The explanation of the EC approach is not clear to me (see my main concern).

Response: We will improve its clarity. Please refer to our response to general comment #2.

Line 94 – 95: Please extend this explanation a bit. It only gets a bit clearer after reading section 3.3. However, as a purely statistical measure, the Spearman's rank correlation does not tell anything about the physical mechanisms behind two variables. Also, it is not clear in this paragraph which "variables" are meant.

Response: We will clarify this and dig deeper about the physical mechanism using other statistical methods such as partial correlations in the revised manuscript.

Line 126 – 130: Please add a critical discussion on the “wet gets wetter” response you find here. There are also several studies that confirm the “wet gets wetter” paradigm only for a small percentage of the land area (e.g. Xiong et al. https://doi.org/10.5194/hess-26-6457-2022), and even Greve et al. (that you cite here) state that “Only 10.8% of the global land area shows a robust ‘dry gets drier, wet gets wetter’ pattern”).

Response: We agree that this “wet gets wetter” signal, primarily identified from global climate models may be obscured when multiple climate drivers and natural variability are considered. Hence, our original conclusions were limited to the agreement of “this ‘wet gets wetter’ signal over water-sufficient lands”. However, evidence from Greve et al. and Xiong et al. shows that differences in study periods, datasets, and the variables used to measure hydrological conditions can substantially influence the inferred responses. Therefore, we will expand the discussion to include arguments both supporting and questioning the “wet gets wetter” response.

Line 135: What does “uniform” storage physics mean?

Response: We don’t assume identical storage physics across different LSMs/GHMs. We will revise this part to clarify this.

Fig 1: I do not understand how mid- and late-century TWS changes can be computed for GRACE/-FO observations (black crosses). This is probably because I did not fully understand the EC approach from the Methods section.

Response: This scatterplot is a conventional way to illustrate the EC relationship. When observations are considered, the focus is typically on the information shown on the x-axis, which is why the mean of observations is indicated by a black vertical line (as it has no corresponding y-axis values). Similarly, the box plots show no information along the x-axis.

Line 176: The ISIMIP2b ensemble contains more and other models that the ISIMIP3b ensemble. Isn’t this the main reason for the different result?

Response: Different suites of models and ensemble members can influence the EC-corrected results, which is why we included as many models as possible. We will acknowledge uncertainties in our results that may arise from differences in the suite of models and ensemble members.

Line 190 – 193: The differences could also be due to different (number of) models used in the ISIMIP3b and ISIMIP2b ensemble, is that correct?

Response: Yes. We will acknowledge uncertainties in our results that may be introduced by the differences in the suite of models and ensemble members at the end of section 3.1.

Line 195: Where does this distinction come from?

Response: We will modify the sentence to improve clarity.

Figure 2 & 3: As far as I understand is Fig. 3d the difference between Fig. 2b and Fig. 3c. However, I do not see the big reduction in northern South America (the big red blob in Fig 2b) being reflected in the difference plot (Fig. 3d). Are you sure these are the correct plots? If so, please comment on this striking pattern in South America in Fig. 2b, and why it is not present in Fig 3c.

Response: Yes, Fig. 3d represents the difference between the two panels (Fig. 2b minus Fig. 3c). We have double-checked the figure, and it is correct. We will also add another figure to the supplementary information. This new figure is nearly identical to Fig. 3, but the EC correction is applied to all grid cells. It is obvious in Fig. S8c that the aforementioned big red blob remains.

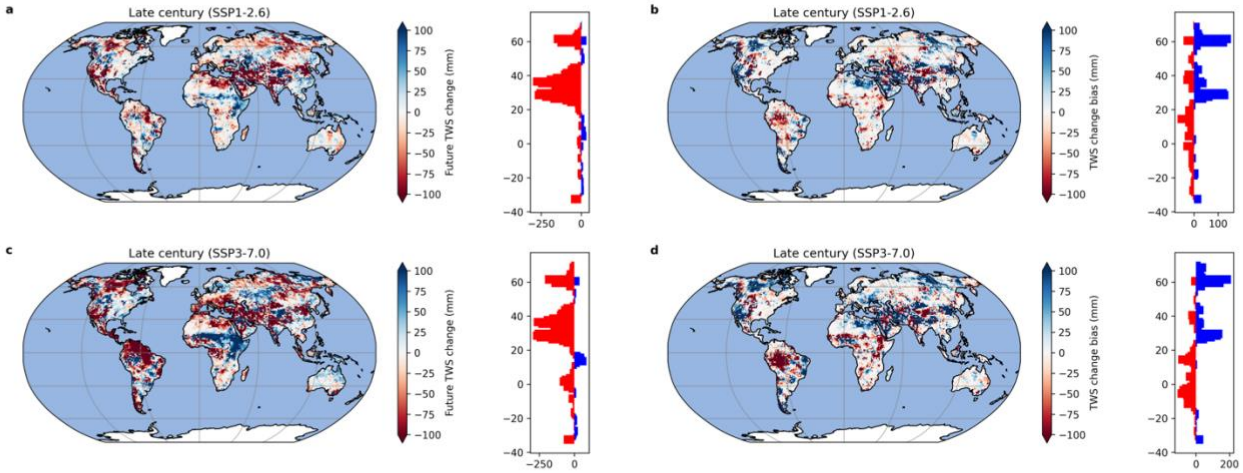

Fig. S8 Same as Figure 3 but the EC correction is applied to all grid cells.

Fig. 3: “only regions with statistically significant positive EC correlations are shown” Please indicate the non-significant regions in another color (e.g. gray) for a better interpretation of the plots.

Response: The colormap used in Fig. 3 is “'RdBu'” (without a white midpoint). Accordingly, all regions shown in white represent grid cells where the results are not statistically significant. We will revise the figure caption to clarify this.

Line 230 – 232: I do not understand in which way Figure 4b and 4c, i.e., the correlation between precipitation and evapotranspiration and runoff, support the findings of the study, or help to better understand physical drivers. In my view you should either extend the analysis considerably or remove evapotranspiration and runoff.

Response: We will expand our analysis in the revised manuscript.

Line 234 – 235: “More importantly, …” I do not understand how this conclusion can be

derived from Figure 4. Please explain it more detailed.

Response: We will conduct partial correlation analysis among different variables to identify true associations between TWS and other variables such as precipitation, evapotranspiration, and runoff.

Fig. 4d: I only see very few black hatches, does this mean all other areas are insignificant? The pattern seems to be quite distinct, therefore I wonder if the significance test might be too pessimistic?

Response: Black hatches indicate statistically significant differences at the 5% level, as determined by Welch’s t-test. A permutation test with 100 random permutations was conducted to estimate the p-values. We also tested a more relaxed threshold ($p < 0.1$), but the significance patterns were identical. We infer that large inter-model variability in TWS among “wet” and “dry” models may lead to statistical non-significance in some regions, even where the differences (wet minus dry models) are substantial.

Line 247 – 251: I find this analysis very interesting, but it should be extended a bit. Where do these differences come from? What is different in “wet” models compared to “dry” models, which processes might lead to this pattern of differences.

Response: We will compare regarding the physical processes and model components of the “wet” with “dry” model groups to provide more insight.

Line 260 – 261: Please explain the interconnection between lack of independent forcing and skewed distribution in Fig. 1 more detailed. For me, it is not straight forward.

Response: Previous studies have applied the EC approach to constrain future projections of the longest annual dry spell (LAD) climatology using CMIP5 and CMIP6 ensembles (Petrova, I.Y., Miralles, D.G., Brient, F. et al. Observation-constrained projections reveal longer-than-expected dry spells. Nature 633, 594–600 (2024). https://doi.org/10.1038/s41586-024-07887-y). That study showed approximately Gaussian distributions in the raw (pre-EC) projections of CMIP5 and CMIP6 ensembles (Fig. 2). In contrast, the skewness observed in the raw projections in our study may stem from the lack of independent forcing across LSMs/GHMs in the ISIMIP framework. However, this interpretation is speculative and requires further supporting evidence. We will therefore revise this section to rely only on more robust results from our analysis.

Line 268: Maybe you can put 83 mm into perspective. As a pure number it does not tell a lot about the significance of the impact the constraining has on the projections.

Response: We will rewrite this section based on the reviewer's comments.

Technical corrections:

Line 60: TWSAs (I think the A was never introduced as abbreviation)

Line 66 and line 194: remove "ref." and brackets

Line 100: typo, resent-day → present-day

Line 107, 115, 124: remove brackets

Fig 1: typo in x-Axis, Hitorical → Historical (and in Fig S1, S2, S7 accordingly).

Fig 3 caption: typo, Figuire → Figure

Line 259: "could also derive" is not a proper sentence

Response: All typos will be corrected in the revised manuscript.