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:

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

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.

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.

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

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

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.

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

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.

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").

Line 135: What does "uniform" storage physics mean?

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.

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

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

Line 195: Where does this distinction come from?

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.

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.

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.

Line 234 – 235: "More importantly, …" I do not understand how this conclusion can be derived from Figure 4. Please explain it more detailed.

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?

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.

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.

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.

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