Revision of the manuscript " Emerging global freshwater challenges unveiled through observation-constrained projections" by Fei Huo, Yanping Li, Zhenhua Li.

General Comment.

In this study, the authors provide an analysis of future terrestrial water storage (TWS) projections using the emergent constraint (EC) methodology. The study employs multiple model ensembles (ISIMIP2b and ISIMIP3b) and GRACE observations and shows that EC calibration reduces uncertainty and corrects biases in raw model projections. The analysis of regional patterns and underlying physical mechanisms strengthens the scientific contribution. However, revisions are needed to clarify statistical methods, explicitly acknowledge uncertainties (model dependence and observational limitations). However, despite my overall positive impression, there are several areas where the paper could be improved, particularly in terms of clarity, methodological rigor, and the interpretation of results. I suggest a major revision.

Response: Thank you for your supportive comments.

My major concerns are listed below.

(1) The concept of emergent constraint (EC) methodology is addressed, but its definition and distinction from the other types of methods could be clarified earlier in the manuscript. They should provide a clear explanation of how this method differs from other methods and why it is particularly relevant in this context.

Response: Emergent constraint is a widely used method to improve climate models' accuracy of future predictions. 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 clarify its strength and the context to apply it in the field of terrestrial water storage (TWS) in the Introduction.

(2) Refining the structure to separate background, problem, and contribution; elaborating on the novelty relative to prior EC studies; and slightly tempering assertive language would substantially improve readability and the significance of the study.

Response: We will refine the content to emphasize the novelty and contribution of our study and use a more assertive tone to better organize the whole text.

(3) Explicitly stating statistical procedures and uncertainty treatments and better articulating the rationale behind certain methodological choices (e.g., scenario selection, regression structure) would substantially strengthen the transparency and reproducibility of the study.

Response: We will add the details of the methodology to clarify the rationale behind the EC method in the Data and methods.

(4) The authors don't address the limitations or uncertainty of the study. The study used GRACE satellite data, while satellite data can have the same biases. In addition, the observed data used to constrain the projection won't be the same in the future. Therefore, you need to state the limitations of the methods.

Response: The limitations of the EC approach and GRACE will be added to the Introduction and the Data and methods.

(5) Doesn't this method depend on the model ensemble structure? Have you looked at the other models?

Response: Yes. A large ensemble size always yields more robust results. Thus, we have included all models provided by the ISIMIP3b and ISIMIP2b projects.

(6) To improve scientific rigor and interpretive balance, I recommend (1) quantifying potential impacts of model dependence, (2) expanding on remaining uncertainties, especially regarding observational and regional limitations. With these revisions, the Discussion would more convincingly convey the robustness and practical significance of the emergent constraint findings.

Response: We will improve the Discussion by expanding on the limitations of the EC method and observations/modelling datasets used in this study.

(7) Several aspects could be improved to enhance clarity, logical structure, and the articulation of the research gap in the introduction section. The introduction covers a wide range of interconnected topics (TWS processes, climate impacts, model uncertainties, EC methodology), but the transitions between them could be smoother. Consider reorganizing into:(1) importance and role of TWS; (2) current challenges and modelling uncertainties; (3) motivation for applying EC and specific research objectives. This will make the narrative flow more naturally from context → problem → solution.

Response: We will organize the Introduction to improve its clarity.

(8) The introduction cites key works Bowman et al., 2018; Brient, 2020; Hall et al., 2019; Petrova et al., 2024; Shiogama et al., 2022) but could better emphasize what has not yet been done. The statement that "potential constraints on global mean changes in terrestrial water storage have yet to be thoroughly explored" is important; consider expanding slightly on why previous EC studies focused mainly on temperature or other hydro-climatic variables, to clarify the novelty and need for this study.

Response: We will rewrite this part to highlight our contribution compared with previous research.

(9) It might also be helpful to briefly mention recent improvements or limitations of GRACE-based datasets (continuity with GRACE-FO), since these are central to the proposed constraint.

Response: The limitations of GRACE and GRACE-FO will be added to the Introduction.

(10) The Methodological Framing in the introduction needs to be clearer. When introducing the EC concept, clarify that it is a statistical relationship across models linking an observable quantity to a future response. This helps readers unfamiliar with ECs understand its basis before applying it to TWS.

Response: We will explicitly let the audience know that the EC approach is basically statistical method to improve future predictions based on connections between historical status and future response.

(11) The final sentence (lines 44-46: "By combining the proposed EC with historical observations from GRACE satellites…") appropriately sets up the study objective but could more explicitly articulate the main research question or hypothesis.

Response: Yes. We will expand on our objective and associated scientific questions in the revised manuscript.

The choice to include all three GRACE mascon solutions (JPL, CSR, GSFC) is well justified. However, please clarify how these were combined, were the datasets were averaged, used separately to estimate uncertainty, or treated as independent realizations? Linear interpolation to fill missing months in GRACE data is acceptable but may introduce temporal bias in regions with strong seasonal variability. Consider noting why linear interpolation was deemed sufficient. It would be helpful to specify whether scaling factors were applied to GRACE data (as is often recommended to correct for signal attenuation), and if not, to justify this choice. The recommendation for Grid Scaling is described here: https://grace.jpl.nasa.gov/data/get-data/monthly-mass-grids-land.

Response: Each GRACE product is treated as an independent realization. Linear interpolation suffices because this study focuses on long-term changes and linear trends in TWS on annual timescales, rather than on resolving seasonal variability. As for the scaling factor, we didn't apply any even though grid-scale factors are recommended for observation-model comparisons, because the objective of our work is to accurately estimate the linear trend in TWS. However, "…, the gain factors tend to be dominated by the annual cycles of land water storage variations, and may thus not be suitable to quantify trends from the GRCTellus land data." (https://grace.jpl.nasa.gov/data/get-data/jpl_global_mascons/). We will improve the pertinent content in the revised manuscript.

The rationale for using both ISIMIP3b and ISIMIP2b datasets is sound, but the selection criteria for the specific LSMs and GHMs should be made explicit. Is it due to completeness, data availability, or model diversity? The decision to merge the end of the HIST run with the beginning of SSP1-2.6 to create a continuous 2004–2023 climatology should be further justified, especially given that this can introduce discontinuities in forcing conditions. Please clarify whether ensemble averaging was performed before regression (to reduce noise) or whether individual model realizations were treated independently.

Response: We included outputs from all available models with overlapping simulations across different scenarios, which is also why we chose SSP1-2.6 to cover the entire evaluation period: most LSMs and GHMs provide outputs under SSP1-2.6. Yes. Using SSP1-2.6 can introduce additional forcing signals compared with HIST, but using SSP2-4.5 would greatly reduce the number of models available for analysis. Each ensemble member is treated as an independent realization. We will revise the manuscript to clarify these points.

The description of the EC implementation is generally clear, but could benefit from additional precision: Specify how the statistical significance of the x–y relationships was assessed. Clarify whether the regression between (x) and (y) was performed globally or per grid cell, as this strongly affects the interpretation and robustness of results. Indicate the sample size (number of models) used in the regression, since ensemble size influences the confidence of emergent relationships. It would be useful to mention whether the linear assumption in the EC regression was tested. When applying the EC to calibrate projections, please specify whether bias correction was applied before regression, and how uncertainty from both observational and model sources was propagated through the EC correction.

Response: We will add the pertinent information on statistical significance, sample size, and the linear assumption to the revised manuscript. Residual plots will be included to check linearity. The regression was conducted globally. Since the purpose of the EC method is to correct model biases using observations, no additional bias correction was applied. Both observational and model spreads influence the future EC correction, with larger spreads indicating greater uncertainty in future projections (i.e., a wider range of outcomes).

The validation step using "six driest and six wettest models" is interesting but could be more systematically described. Were these classifications based on percentiles of TWS climatology, or another quantitative criterion? Also, it might be useful to test whether "dry" and "wet" subsets produce statistically distinct EC relationships.

Response: The "dry" and "wet" models were categorized by ranking the absolute values of their TWS climatology, which is similar to a percentile-based criterion. We will use different criteria (for example, percentiles of TWS climatology) for classifications instead to test the sensitivity of the results. Furthermore, we will add the EC results obtained only using the "dry" or only the "wet" model subsets to the supplementary information, where interested readers can examine them in detail.

I have concerns regarding the statistical Strength and Presentation of the EC Relationship. The reported "Significant positive correlations ($R > 0.99$ for both ISIMIP2b and ISIMIP3b models) are found between historical and late century annual area-weighted global mean TWSAs, irrespective of the emissions scenario" appear high. Please clarify whether this reflects ensemble-mean correlations (i.e., correlation of means across models) or model-level correlations across realizations. Indicate whether statistical significance was adjusted for the number of ensemble members or grid points.

Response: Our results are consistent with previous research, which found a high correlation coefficient ($R = 0.98$) between historical and future climatology of the longest annual dry spell (LAD) based on CMIP5 and CMIP6 ensembles (Extended Data Fig.4, 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). Such a high correlation in LAD, a key drought indicator, implies a global "dry-model-gets-drier" relationship in climate model simulations. This relationship can propagate into LSMs and GHMs through climate forcing. R was calculated using model-level realizations instead of ensemble means to increase the sample size. Statistical significance of correlations was adjusted for the number of ensemble members (i.e., 25 for ISIMIP3b and 31 for ISIMIP2b).

The authors note the use of both ISIMIP2b and ISIMIP3b ensembles, but it remains unclear whether the emergent relationship was derived separately for each dataset or jointly across all models. Explicitly stating this would improve transparency.

Response: The EC corrections were applied separately to the ISIMIP2b and ISIMIP3b datasets.

The authors state, “Furthermore, the EC correction constrains the discrepancies of late century TWS changes by 63% for the SSP1-2.6 scenario and 69% for the SSP3-7.0 scenario”. This reduction should be quantitatively defined if it decreases in variance, range, or standard deviation. and accompanied by confidence intervals or bootstrap uncertainty ranges.

Response: We will revise the manuscript to examine TWS changes in more detail.

The Interpretation and Physical Plausibility of the physical explanation are not clear. The author states that the results are consistent with the “wet gets wetter” paradigm, but the physical explanation could be expanded to discuss exceptions, regions, or model subsets where this relationship does not hold.

Response: We will expand our analysis in the mechanism part, especially focusing on regions where this paradigm does not hold.

The statement, in lines 169-170, that EC-corrected results “produce more robust projections than conventional approaches,” is reasonable but somewhat strong. The authors acknowledge in lines 172-174 that the discrepancy arises because GRACE products used for EC correction are all derived from a single source, the GRACE satellites, resulting in an underestimated uncertainty for global mean TWS. This is an important caveat and should be emphasized more explicitly as a potential limitation rather than a parenthetical remark.

Response: The limitations will be acknowledged explicitly in the revised version.

Need to describe Figure 2 clearly. Line 186, you describe Figure 2a as the TWSAs, but in line 189, you use “TWS changes”. But in the figure caption, you use “ensemble averages of TWS changes”. Figures 2 and 3 are central to the paper, but the text and the caption describing them could be clearer. Indicate explicitly how the EC calibration modifies the spatial distribution. The finding that only ~26% of land areas show significant EC correlations ($R > 0$, $p < 0.05$) is interesting but somewhat low; consider discussing why many regions do not exhibit a significant relationship.

Response: We will improve the clarity of the captions for Figures 2 and 3. The discussion about how the EC correction changes the spatial pattern will also be added. Although the fraction of land areas with significant EC correlations is ~26%, this value is comparable to that reported in a previous study, which used the EC calibration on the longest annual dry spell (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). This limited spatial extent likely reflects the complexity of the hydrological response to global warming. Despite the globally identified "wet gets wetter" atmospheric response in general circulation models (GCMs), the terrestrial water cycle is governed by more intricate mechanisms. In many regions, freshwater variability is strongly influenced by natural variability, including oscillations between dry and wet periods driven by El Niño, La Niña and other climate modes.

The overestimation of TWSA in northern midlatitudes and underestimation in humid regions aligns with prior findings, but it would be useful to quantify regional biases. The correlation between precipitation and TWSA changes (Fig. 4) supports the physical validity of the EC, but the analysis might benefit from partial correlation tests to control for precipitation when relating TWSA to other drivers (e.g., evapotranspiration or runoff).

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

The "wet vs. dry model" comparison is intriguing, but the classification criteria are somewhat arbitrary; consider clarifying or showing that results are insensitive to the chosen threshold. In discussing the "wet gets wetter" mechanism, it would be useful to acknowledge nonlinear land surface feedback (groundwater depletion, vegetation response) that may dampen or reverse this pattern regionally.

Response: We will use different criteria (for example, percentiles of TWS climatology) for classifications instead to test the sensitivity of the results. The nonlinear interactions among different land components will also be acknowledged.

Several areas require further clarification or refinement to improve balance, transparency, and interpretive rigor in the discussion section. The authors identify that the reliability of proposed ECs could be compromised due to the lack of independence among climate models (Brient, 2020; Caldwell et al., 2014). However, the discussion could be more quantitative. The statement that "diversity in global models and their climate forcings is critical" is well-taken but could be strengthened by suggesting specific strategies, for instance, promoting structural independence in ISIMIP protocols or incorporating multi-forcing experiments to test robustness.

Response: The discussion will be refined to be more interpretive.

The reported “average TWS decrease of roughly 83 mm” is a key quantitative finding, but should be contextualized: what fraction of total terrestrial water storage does this represent? And how does this compare to previous assessments? The phrase “elevates the risk of basins being underprepared” introduces a policy implication that is not directly analyzed in the study. Consider softening this to “may imply that current water resource planning could underestimate potential shortages,” unless explicit basin-level analyses were performed.

Response: We will improve this part based on the reviewer’s comment.

The conclusion could be better with the limitations and broader implications. While model dependence and structural similarity are mentioned, other potential limitations deserve brief acknowledgment, such as short observational baselines, GRACE uncertainty, and non-stationary relationships under extreme forcing scenarios. The conclusion could better distinguish between confidence in the global-scale EC relationship (which appears robust) and uncertainty in regional-scale projections, where the EC significance covers only ~26% of land areas. This nuance would enhance interpretive caution. The section might benefit from a short paragraph linking the EC findings to future modelling priorities.

Response: Many thanks. We will expand the discussion of limitations from broader perspectives and analyze the uncertainty on both global and regional scales. Also, future modelling priorities will be outlined in the concluding section.

Specific comments:

The abstract mentions “low- and high-end forcing scenarios” without naming them explicitly (e.g., SSP1-2.6 and SSP3-7.0), which would aid clarity and reproducibility.

Response: The information will be added to the abstract.

Line 28-31: It is not always the case, as the warming climate can modify precipitation patterns and lead to floods in some regions. So, saying that “these changes exacerbate freshwater scarcity” is not always true for all regions.

Response: We will rephrase the text to ensure an accurate statement.

Line 38: What is the motivation for using the “emergent constraint (EC) approach”

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.

Line 45: The phrase “successfully constrain future TWS changes could be rephrased to avoid implying confirmed success, e.g., “apply the EC framework to constrain projections of future TWS changes.”

Response: Modified as suggested.

Check that reference years (e.g., GRACE citations) correspond to the latest data versions.

Response: Modified as suggested.

Line 49: Why did you choose the period 2004-2023? Why not consider a long period from [Humphrey, V., & Gudmundsson, L. (2019). GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. Earth System Science Data, 11(3), 1153-1170.] https://figshare.com/articles/dataset/GRACE-REC_A_reconstruction_of_climate-driven_water_storage_changes_over_the_last_century/7670849

Response: GRACE-REC is a widely used reconstructed dataset, but as recommended by the authors of Humphrey, V., & Gudmundsson, L. (2019), “the reconstructed TWS trends mainly depend on the trends initially present in the driving precipitation data”, and “it should be clear that there will be differences between the trends found in GRACE and the trends found in the reconstruction. Such discrepancies are expected because the reconstruction does not represent several sources of long-term changes in TWS…”. Therefore, we didn’t use GRACE-REC in our study because capturing linear trends are crucial to our analysis.

Line 49-57: The authors don’t state the name of the variable used in the study in all paragraphs of the Observation section, which is important for the readers.

Response: Modified as suggested.

Line 56-57: Give more explanation regarding the linear interpolation model. The authors need to convince me that this approach is acceptable. Why not consider the Humphrey, V., & Gudmundsson, L. (2019), which is a reconstruction of data.

Response: Our study focuses on long-term changes and linear trends in TWS on annual timescales, rather than on resolving seasonal variability. As we mentioned above, GRACE-REC is more suitable for analysis of interannual variability. The author of Humphrey, V., & Gudmundsson, L. (2019) stated, and I quote: “the trends in GRACE-REC cannot be directly evaluated against the trends from GRACE itself.”

Line 62: What are those five general circulation models (GCMs)?

Response: We will improve clarity in the revised manuscript.

Line 63: Ensure consistent notation: SSP should be defined at first use in the main text.

Response: Modified as suggested.

Line 73: Which method have you used to regrid the data?

Response: Details added.

Line 74: Why was the comparison only made relative to the period 2004-2009? Is it no too short?

Response: The 2004-2009 baseline time was used solely to calculate monthly climatologies and corresponding anomalies. As long as this baseline (climatology) is fixed (i.e., constant at a given grid point), its short duration doesn’t affect the results because our focus is on long-term changes and linear trends of these anomalies. We will clarify this in the revised version.

Line 100: What does “resent-day global mean TWSAs”? Is it a typo: “resent-day” → “present-day.”

Response: Modified as suggested.

Lines 106-107: Ensure consistent reference formatting and in-text citation style (e.g., “(for example, (Cai et al., 2025; …)” should remove nested parentheses).

Response: Modified as suggested.

Line 168-170: Phrases such as “produce more robust projections” and “further exacerbating existing water stress worldwide” are somewhat strong. Consider tempering the language to reflect the uncertainty and limitations of observational constraints.

Response: Modified as suggested.

-Fig S1 caption: In the legend of Fig S1a, you describe the crosses for observation, but you don’t state it in the Fig S1 caption, “Dots and crosses represent global averages of TWSAs from ensemble members.” As same as for FigS2.

Response: Modified as suggested.

It will be convenient to add the observation in Fig. S1.b to make it for comparison. -It will be better to have a unique color bar for the figS4 to represent the difference. For (a) and (b), you used red to represent the decrease in TWS and precipitation, while it is blue for the Evapotranspiration and Total runoff. The same for FigS5, FigS6.

Response: Modified as suggested.

-There is only 5 crosses from observation in the Fig. S7, while you declare that you replaced the mascon solutions with 7 Grace-derived TWSA.

Response: Some datasets show nearly identical values (e.g., -13.10 for JPLM and -13.72 for GSFC; -9.61 for JPL and -9.47 for GFZ), causing their crosses to overlap in the figure.

- Be consistent with the format of supplementary figure citation in the main manuscript, for example, in line 179, you use “Supplementary Fig. 7” but in line 180 it is “Fig. S2”. And make Sure that all supplementary figures are addressed in the main manuscript.

Response: Modified as suggested.