Editor decision: Reconsider after major revisions (further review by editor and referees)

by Rohini Kumar

Public justification (visible to the public if the article is accepted and published):

Dear Authors,

Thank you for actively participating in the discussion round and posting your responses to the referees’ comments. I appreciate your detailed and constructive responses to the referee's comments. Besides the modification you proposed in response to the referee's suggestions; I would like to point out that the suggestion of referee #1 regarding further deepening into other water fluxes and storages (ET, Q, SM) is legitimate and needs to be addressed in the revision for "better understanding of the reasons why GRACE and models differ from each other". Also wherever you claim that your adopted approach is standard and commonly used (e.g., linear interpolation or use of JPL-M and CSR-M GRACE solutions) - try to justify them with proper examples/references/citations, etc (and not just by words). I also see the point of reviewer #3 on using the mean of two GRACE products instead of relying on just one GRACE product to evaluate models; or also using the third GRACE JPL-M solution which so far is missing from your analysis. These are good points raised by the referees and you should consider them in your revision.

In your revision, please provide a point-to-point answer to the comments made by the referees along with a track-changed version of the revised manuscript. Since revising your manuscript following the referee's comments requires some major changes, after receiving your revised manuscript I will send the revised work out to the referees for a second round of reviews.

I look forward to receiving your revision.

Best regards,

Rohini Kumar

Reply to Editor’s Decision Letter

Dear Dr. Kumar,

Thank you for your valuable feedback on our manuscript. We appreciate the thorough review and constructive comments provided by the referees. We are pleased to hear that you found our responses detailed and constructive.

In the revised version of our manuscript, we have addressed the suggestions put forth by referee #1 regarding a more in-depth exploration of other water fluxes and storages, including evapotranspiration (ET), streamflow (Q), and soil moisture (SM). We have incorporated additional figures and discussions on these components, aiming to provide a comprehensive understanding of the reasons for discrepancies between GRACE and models.

Furthermore, in response to the suggestion to justify our adopted approaches with proper examples/references/citations, we have strengthened our manuscript by incorporating relevant references to support our use of linear interpolation and the choice of JPL-M and CSR-M GRACE solutions. We understand the importance of substantiating our claims with concrete evidence and have taken this into consideration in our revision.

Regarding the recommendation from reviewer #3 to use the mean of two GRACE products for model evaluation, we have implemented this suggestion in the revised version and we used the mean of two GRACE solutions. These modifications enhance the robustness of our study and contribute to a more thorough evaluation of the models.

We believe that these revisions address the concerns raised by the referees and strengthen the overall quality of our manuscript. Thank you once again for the opportunity to revise our manuscript, and we look forward to your further guidance in the next steps of the review process.

Sincerely,

Tingju Zhu

Response to Comments of Referee #1

Comment: The comparison is mainly implemented for different climate zones. However, as we known, GWS may dominate TWS in many regions, thus the comparison between the models and GRACE is better to be divided into the models with and without groundwater simulations.

Reply: Thank you for your valuable feedback. We appreciate your input and have taken your suggestion into consideration. As you rightly pointed out, groundwater storage (GWS) can indeed dominate total water storage (TWS) in many regions, like basins which can significantly impact the comparison between model simulations and GRACE data.

In response to your suggestion, we have divided our comparison over five river basins with major underlying aquifers (Congo, Amazon, Orinoco, Ganges-Brahmaputra, and California) into models with and without groundwater simulations. Please see Figure 12. The comparison suggests that incorporating a groundwater compartment in GHM/LSM models can enhance the representation of water storage dynamics in certain basins, although this improvement may not be applicable across all basins.

Comment: More investigations on the water cycle (e.g., P, ET and Q) and storage (e.g., GWS, SMS) compartments will be helpful for a better understanding on the reasons why GRACE and models differ from each other in the aspects of phase and amplitude.

Reply: Thank you for your insightful comment. We appreciate your suggestion for more investigations on the water cycle and storage compartments. In our revised version, we have incorporated additional figures (13-15) that specifically address total soil moisture, ET, and total runoff. These additions aim to provide a more comprehensive understanding of the factors contributing to the discrepancies between GRACE and models in terms of phase and amplitude.

Comment: Figure 1-4: what the map colors mean?

Reply: We provided the description of base map in captions of Figures 1-4. Base map represents KGClim Climate Zones classification, and additional details can be found in Figure S2 within the supplementary material.

Comment: Table 1: Statistical information on the phase and amplitude derived from models and GRACE needs to be provided as they are the key information. It can be included in Table 1 or summarized in another table.

Reply: The statistical analysis on amplitude derived from models and GRACE is provided in Table 4 and phase analysis is provided in Figure 5.

Comment: Line 83: Full names are need for the abbreviation for R1 and R2 at its first time.

Reply: Full names of R1 and R2 are provided; please see the lines 24-26.

Response to Comments of Referee #2

This study compares the performance of 13 GHMs and LSMs in capturing amplitude and phase of TWSC in global major rivers against GRACE data, including comparisons across climate zones and model version (R1 and R2). This detailed comparison facilitates improved parameterization model process. However, limitations of this study include overly detailed descriptions of the comparisons of different basins so that it is difficult for the reader to get to the point, and the figure lacks summarization.

Major comments:

1. The configuration of the modules of each model should be clearly stated, and some differences may be due to missing modules, e.g., snow, permafrost, groundwater, etc., which would also be useful for analyzing the causes of deviations. And, if key modules are missing does it still make sense to compare changes in TWSC between the model (GHMs and LSMs) and GRACE.

   Response: Thank you for your valuable feedback on our manuscript. We appreciate your thoughtful comments and suggestions. In response to your specific points:

   We agree that providing a clear and comprehensive description of the configuration of the modules in each model is essential. In our revised manuscript, we included details of key modules within each model in Table 2, including any specific modules for snow, groundwater, and other relevant components. This will help readers understand the differences and similarities between the models and their potential impact on the results. We acknowledge the importance of considering the potential impact of missing key modules in our models. While some differences between the models and GRACE may indeed be due to these missing modules, we believe that the comparison still holds value. We aim to assess the agreement and discrepancies between the models and GRACE in terms of Total Water Storage Changes (TWSC) to better understand the limitations of both approaches. By highlighting the missing modules in section 4.1, we provided insights into the potential sources of deviations and uncertainties in TWSC estimates.

1. This study did not analyze in depth the causes of amplitude and phase differences, especially 4.1 section.

   Response: Thank you for your valuable feedback. We appreciate your suggestion to delve deeper into the causes of amplitude and phase differences in Section 4.1. In response to this comment, we expanded the discussion in Section 4.1 and provided a more comprehensive analysis of the factors contributing to the observed amplitude and phase differences.

2. Line67, “due to human intervention and climate change respectively”, the underestimation is due to anthropogenic interventions and climate change, doesn't that have anything to do with model performance, shouldn't model performance be the main reason?

Response: Thank you for your comment. We corrected lines 70-71.

3. Line 89, amplitude, and phase of "polar" zone was not analyzed in result section.

   Response: Thank you for your valuable comment. In this study, we primarily concentrated on analyzing the boreal, temperate, arid, and tropical zones, we did not include the polar zone in our analysis. However, we believe that exploring the amplitude and phase of the polar zone could indeed be a valuable avenue for future research to provide a more comprehensive understanding of the subject matter. We will duly consider this suggestion for future studies in this field.

4. Line 139, Why not CSR and JPL on average?

   Response: Thank you for your observation. In the revised manuscript, we opted to use the average of CSR and JPL data for greater representativeness and a more comprehensive analysis.

5. The difference in the length of the text in parts 3.1 and 3.2 is too large. 3.1 section over-emphasis on basin comparisons.

   Response: Thank you for highlighting the difference in text length between sections 3.1 and 3.2. We acknowledge your concern. The length of section 3.1 was extended to appropriately address the variations in amplitude and peak magnitude observed across different basins. However, we understand the need for a balanced presentation. To address this, we have revised and streamlined section 3.1 to maintain focus on basin comparisons without unnecessary elaboration. However, achieving equal length in sections 3.1 and 3.2 may not be feasible due to the inherent differences in the nature of the data and analysis. We have, yet, made efforts to ensure that both sections maintain a proportional and justified length based on the complexity and variability present in each aspect of the study. Your input was valuable, and it improved the overall flow and readability of our manuscript.

6. The figures are not summarizing enough, too many similar comparisons, e.g., I think Figures 5-8 should be in the Appendix, and the main results should be put in the main text, e.g., the overall results for the different climatic zones in one fig.

   Response: Thank you for the reviewer's comment regarding the figures in our manuscript. We understand your concern about the number of comparisons and the desire for a more concise summarization. However, we believe that Figures 6-9 (in revised manuscript) are important for understanding the detailed results and patterns in different regions and should remain in the main text.

7. I suggest to add the spatial distribution map of biases in amplitude and phase.

   Response: Thank you for your suggestion to include spatial distribution maps of biases in amplitude and phase. We understand the importance of visualizing these biases for a comprehensive understanding of the results. However, we want to clarify that such maps have already been provided by Schellekens et al. (2017), and our study relies on their analysis in this regard. Including redundant maps in our paper would indeed be repetitive and not add significant new insights to the existing literature.

We appreciate your concern, and to ensure clarity in our paper, we will explicitly reference and acknowledge the work of Schellekens et al. (2017) for the spatial distribution maps of biases in amplitude and phase between the models and GRACE data. This will help readers access the relevant information in the cited source while maintaining the focus of our study on its unique contributions and analyses.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P.: A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset, Earth Syst Sci Data, 9, 389–413, https://doi.org/10.5194/essd-9-389-2017, 2017.

8. Figure 1-4 suggests the addition of lines for the GHM and LSM model averages, which facilitates comparison of the two types of models

   Response: Thank you for your feedback regarding Figure 1-4. We appreciate your suggestion to add lines for the GHM and LSM model averages to facilitate a clearer comparison between the two types of models. We have now incorporated these lines into the figures as per your recommendation. This enhancement should provide readers with a more comprehensive view of the model comparisons and improve the overall clarity of the presentation.

Minor comments:

1. Line 4, "(e.g., the amount and" misses the corresponding right parentheses.

   Response: We appreciate your suggestion and added parenthesis in the revised manuscript, please see line 38.

2. Line68, "Other studies focused on the seasonal cycle of TWSC to identify" to "Other studies on the seasonal cycle of TWSC focus on identifying" is more suitable? "disparities", specifically what are the disparities?

   Response: Thank you for your suggestion. We have revised the sentence as follows: "Other studies on the seasonal cycle of TWSC, such as Zhang et al. (2017), have focused on identifying disparities." The term "disparities" refers to differences or variations in four global numerical model realizations that simulate the continental branch of the global water cycle and GRACE that have been investigated in previous studies.

3. Line 75, "northern basins" is vague, please specifically point

   Response: Thank you for the suggestion. We have made the requested clarification in the manuscript. Line 76-77 now reads, "northern high-altitude basins," to provide a more specific description of the geographic region being referred to. This should help eliminate any ambiguity and ensure a clearer understanding for the readers.

4. Line 84, “replicate water storage against the latest release (RL06) of GRACE TWSC.”, this sentence indicates the result? this place is to say what is to be studied

   Response: We appreciate your feedback. We rephrase the sentence to “Compare high-resolution and more optimized structured R2 models against R1 models and access their ability to simulate TWSC variability and replicate water storage against GRACE TWSC.

Response to Comments of Referee #3

Bibi et al. evaluated the reliability of 13 global models using the GRACE TWS for 29 river basins. They conclude that the modeled TWS does not compare well with the GRACE TWS. Authors analyzed amplitude and phase-difference and performed the comparisons based on 5 climate zones (Polar, boreal, temperate, arid, and tropical), 2 set of hydrological model types (LSM and GHM), and 2 sets of R1 and R2 model types. The authors find that R2 models have better correlations with GRACE than R1 models. Though this study provides new insights into the future improvement of large-scale hydrological models, there are some major concerns in this study. By addressing these concerns, the manuscript will better align with the standards of the HESS and provide a more compelling and novel contribution to the field. Please find my detailed review below-

Comment: Line 22- It would be easier for readers to understand if the meaning of the term 'R2 models' is provided here.

R1 and R2 models are Water Resource Reanalysis tier-1 and tier-2 products which provide a large set of LSMs and GHMs.

R1: 0.5° forced with ERA-Interim data (WFDEI) meteorological reanalysis dataset

R2: 0.25° forced with Multi-Source Weighted Ensemble Precipitation (MSWEP) dataset

Response: Thank you for your valuable comment regarding the terminology used in our manuscript. We appreciate your suggestion to clarify the meaning of 'R1 and R2' models' for the benefit of our readers. In response to your comment, we have added a brief explanation of the terms in the manuscript to improve clarity, please see lines 24-26. We hope that this addition will enhance the understanding of our work for all readers.

Comment: Lines 91-102: The authors have used only JPL-M and CSR-M solutions, why not GSFC Mascons as well? The authors did not provide the reason behind using linear interpolation of GRACE TWS data.

Response: Thank you for your valuable feedback, we appreciate your input and would like to address your comments:

In our study, we used the GRACE JPL-M and CSR-M solutions for several reasons. These two solutions are widely recognized and have been extensively validated in the literature i.e., (Schellekens et al., 2017; Scanlon et al., 2021) and are among the most commonly used GRACE solutions for global terrestrial water storage (TWS) estimates due to their accuracy. However, we acknowledge that incorporating the GSFC Mascons solution in future research could provide additional insights, and we will consider this for future work.

We used linear interpolation as it is the most commonly used methods to fill the gap in GRACE record i.e., (Xiao et al., 2015; Liesch & Ohmer, 2016), it is computationally efficient and straight forward and preserve linear trends in the data.

Thank you for your thoughtful comments, and we provided references and explanation in our revised manuscript, please see lines 109-113.

Comment: Lines 109-110: Please correct the sentence.

Response: Thank you for pointing out the issue with the sentence in lines 109-110 of our manuscript. We apologize for any confusion. We corrected the sentence to ensure it is clear and accurate in the revised manuscript, please see lines 130-135. Your feedback is greatly appreciated.

Comment: Why only amplitude and phase of seasonal cycle of TWS was checked in this study? Why not the trend in the TWS data?

Response: Thank you for your valuable comment regarding the analysis of GRACE TWS data in our study. We appreciate your feedback, and we'd like to respond to your question:

In our study, we focused on analyzing the amplitude and phase of the seasonal cycle of TWS for several reasons. The primary objective of our research was to assess the seasonal variability of terrestrial water storage (TWS) in a specific region. Seasonal changes in TWS are of significant importance for various applications, such as hydrological modeling, agriculture, and water resource management. Therefore, our study aimed to understand how well the GRACE and models captured these seasonal variations. The amplitude and phase of the seasonal cycle provide crucial information about the timing and magnitude of TWS changes, which are particularly relevant for addressing certain research questions.

Our research objectives were specifically tailored to examine the seasonal patterns of TWS in the study area during a particular time frame. While assessing trends in TWS is indeed important for different research questions, it may require a separate analysis and may involve addressing different objectives. We decided to focus on the seasonal cycle for the sake of clarity and to maintain a concise scope within the context of our study.

However, we acknowledge that the analysis of trends in TWS data is a valuable avenue of research, and it can provide insights into long-term hydrological changes. We hope this explanation clarifies our choice to focus on the amplitude and phase of the seasonal cycle of TWS in this particular study. If you have any further questions or suggestions, please feel free to let us know. Your feedback is greatly appreciated.

Lines 138-139: Why only GRACE CSR_M seasonal cycle was used to validate the model results? As indicated above why GRACE JPL-M data was not used? Or the mean of the two datasets?

Response: Thank you for your insightful comment. In the revised manuscript, we have taken your suggestion into account and incorporated both CSR-M and JPL-M data to validate our model results. By using the average of these datasets, we aim to provide a more comprehensive validation approach, considering the strengths and characteristics of both CSR-M and JPL-M solutions, please see line 142.

Comment: Line 375-: The causes of discrepancies in seasonal amplitudes and phase between models and GRACE TWSC provided in section 4.1 are without any reference. There is no analysis shown to backup the claim. For example, how do the authors know that Model Parameterization is causing the difference in GRACE and model TWS data without doing any analysis and citing any literature? If it is well known then what is the contribution of this study?

Response: Thank you for your valuable feedback. We appreciate your suggestion to delve deeper into the causes of amplitude and phase differences in Section 4.1. In response to this comment, we expanded the discussion in Section 4.1 and provided a more comprehensive analysis of the factors contributing to the observed amplitude and phase differences with reference.

Comment: Scanlon et al., (2018) already compared the model TWS trends against the GRACE TWS datasets. What are the novel contributions here? Please state them clearly.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., Van Beek, L. P., ... & Bierkens, M. F. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. Proceedings of the National Academy of Sciences, 115(6), E1080-E1089.

Response: Thank you for your thoughtful review comment. We appreciate your engagement with our work and the opportunity to clarify the novel contributions of our study in "Benchmarking multimodel terrestrial water storage (TWS) seasonal cycles against GRACE observations over major global river basins", especially in light of the previous work by Scanlon et al. (2018).

While it is true that Scanlon et al. (2018) compared model TWS trends against GRACE TWS datasets, our study focuses on a distinct aspect of TWS analysis, namely, the seasonal cycle. We acknowledge the prior work of Scanlon et al., (2018), which primarily delved into decadal trends in TWS and highlighted discrepancies between models and GRACE observations over extended time periods. In contrast, our study shifts the focus to the seasonal variations in TWS, with the following novel contributions:

We specifically investigate the seasonal dynamics of TWS across major global river basins. Instead of examining long-term trends, our study provides a detailed examination of how TWS varies throughout the year. The phase difference between GRACE and the modeled TWS seasonal cycle was not generally covered in previous studies.

We employ 13 models, each with its own set of assumptions and parameters, to assess how well they capture the observed seasonal TWS variations. This approach enables us to assess the performance of different models in representing seasonal patterns, which can have important implications for water resource management, flood forecasting, and ecosystem health.

While Scanlon et al. (2018) did use GRACE data as a reference, our study explicitly benchmarks the seasonal TWS cycles produced by various hydrological models against GRACE observations. By doing so, we assess how well these models capture the seasonal dynamics observed from space, which can reveal model strengths and weaknesses in representing short-term hydrological processes.

In summary, our study takes a different angle in the assessment of TWS by focusing on seasonal variations and conducting a comprehensive benchmarking exercise using multiple models against GRACE observations. This approach offers valuable insights into the performance of hydrological models in simulating short-term TWS dynamics, providing critical information for applications such as water resource management, drought monitoring, and flood prediction.