Cover Letter

October 16, 2023

Dear Editor,

We would like to thank you and the anonymous reviewers for constructive comments and suggestions, which have significantly improved our manuscript (essd-2023-315).

Terrestrial water storage (TWS) includes all forms of water stored on and below the land surface, and is a key determinant of global water and energy budgets. However, TWS data from measurements by the Gravity Recovery and Climate Experiment (GRACE) satellite mission are only available from 2002, limiting global and regional understanding of the long-term trends and variabilities in the terrestrial water cycle under climate change.

To our knowledge, this study is the first to reconstruct TWS extending to 1940 at a spatial resolution of $0.25°$ at a global scale. Along with its extensive attributes, GTWS_MLrec can support a wide range of geoscience applications such as better understanding the global water budget, constraining and evaluating hydrological models, climate-carbon coupling, and water resources management.

In this revision, all the reviewers' concerns have been addressed. Changes made in the revised manuscript are coloured in blue. It would be greatly appreciated if the revised version of the paper could be re-evaluated by the same reviewer who spent considerable time to provide constructive and professional comments and suggestions, which have led to significant improvement of the presentation and quality of the paper.

We sincerely hope you will find the revised version of the paper appropriate for publication. All authors have reviewed the paper and agree to the resubmission of the manuscript. We look forward to hearing from you.

Sincerely yours,

Dr. Jiabo Yin, Associate Professor
State Key Laboratory of Water R & H Engineering Science
Wuhan University, China
jboyn@whu.edu.cn

Reply to Reviewers' comments

Legend

Reviewers' comments

Authors' responses

Direct quotes from the revised manuscript

Reviewer #1

This study introduces an extended and detailed dataset of terrestrial water storage (TWS) anomalies covering the period from 1940 to 2022 with a spatial resolution of 0.25 degrees. The dataset, named GTWS-MLrec, was generated using machine learning models that incorporate various predictors, including climate, hydrology, land use, and vegetation data. GTWS-MLrec seems to align well with GRACE/GRACE-FO and with other hydroclimatic variables, and seems to outperform previous TWS datasets in reliability. The dataset includes multiple reconstructions based on different mascon sources and provides detrended and de-seasonalized versions. It also covers global average TWS for land areas. The paper is also well-written and the publicly accessible GTWS-MLrec hereby seems to be a valuable resource for various geoscience applications. Therfore, I recommend the paper being published essentially as is (see extremely minor comments below):

Reply: We appreciate the reviewer's positive evaluation of our manuscript and the insightful comments on the improvement of the manuscript. All the concerns have been addressed in the revised manuscript; we hope the revised manuscript satisfies these concerns for publication!

I understand the processing is complex, and very computationally demanding but is there any code available (upon request?) in case other authors want to reproduce (or amend) any of the methods.

Reply: We will upload our codes in the final dataset link with the final version. The current dataset link is https://zenodo.org/record/8187432.

Why is the relationship between streamflow and TWS evaluated using pearson correlation coefficients? I could imagine this relationship being nonlinear?

Reply: The relationship between streamflow and TWS might be non-linear, so we will provide more discussions in the *Section 6 Summary, applications, and outlook* as follows:

We have evaluated the relationship between our reconstructed TWS and streamflow measurements at 10,168 gauges in terms of PCC. TWS and streamflow might have a non-linear relationship due to their different generating mechanisms. However, it is difficult to quantify their physical non-linear relationship due to observational data limitations. Previous studies also evaluated the performance of TWS reconstructions by exploring their linear relationship with streamflow (e.g., Humphrey and Gudmundsson 2019; Li et al., 2021, 2022). In future works, a more complicated non-linear relationship between TWS and streamflow might further evaluate the performance of TWS reconstructions.

Is there any way to make Figure 4 better readable?

Reply: Thank you; we will improve the readability of Figure 4 by changing line colors and types as follows. We also find that the publication form (.pdf file) of this figure is clear.

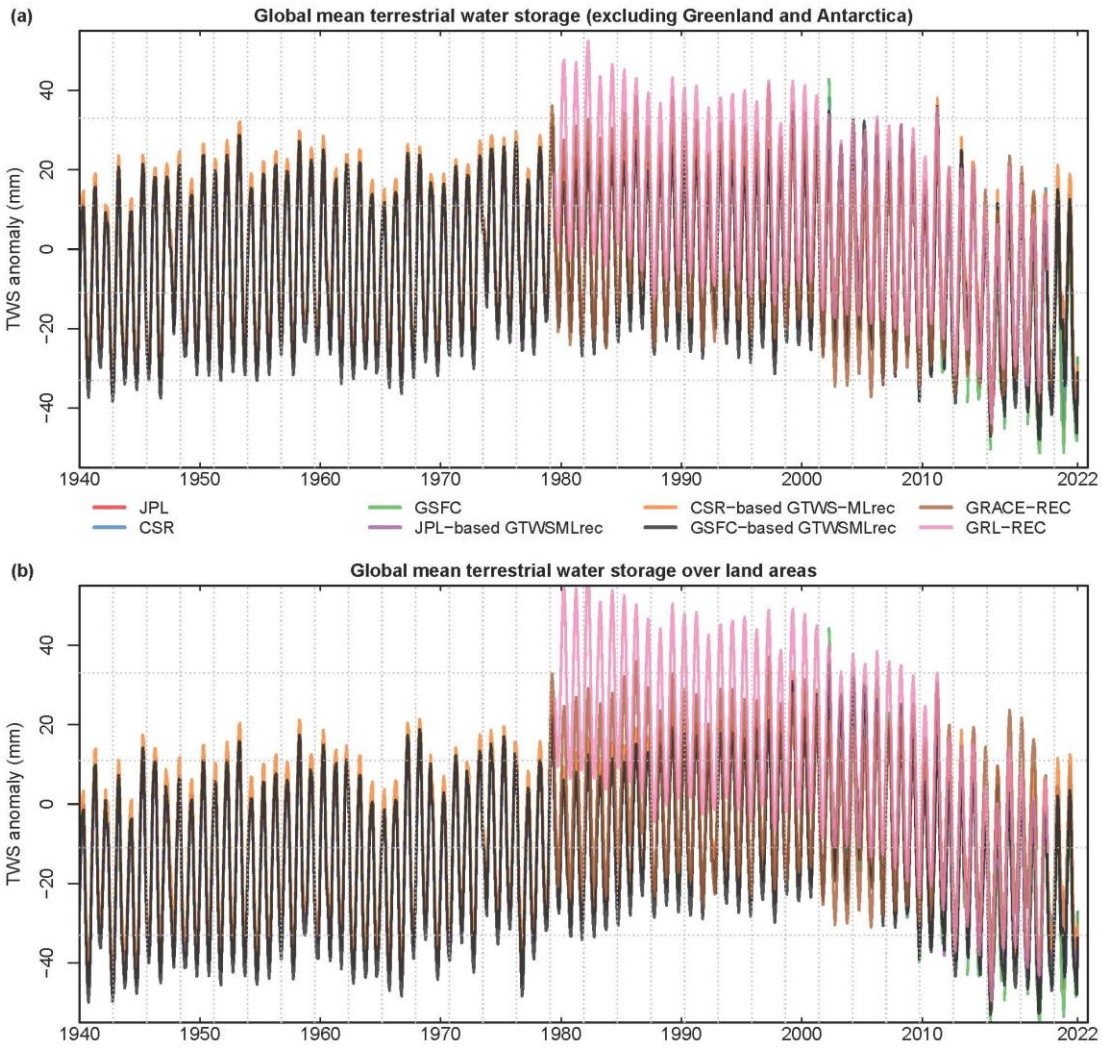

Figure 4. Global mean monthly terrestrial water storage anomaly derived by eight different datasets (including GRACE/GRACE-FO observations). (a) Global average TWS anomaly weighted by land area excluding Greenland and Antarctica; (b) Global average TWS anomaly over land areas.

Figure 8: there is a lot of overlap in markers on the map which makes it sometimes hard to interpret.

Reply: We will shrink the station points and also use transparent colors to make Figure 8 more interpretable. We also find that the publication form (.pdf file) of this figure is clear.

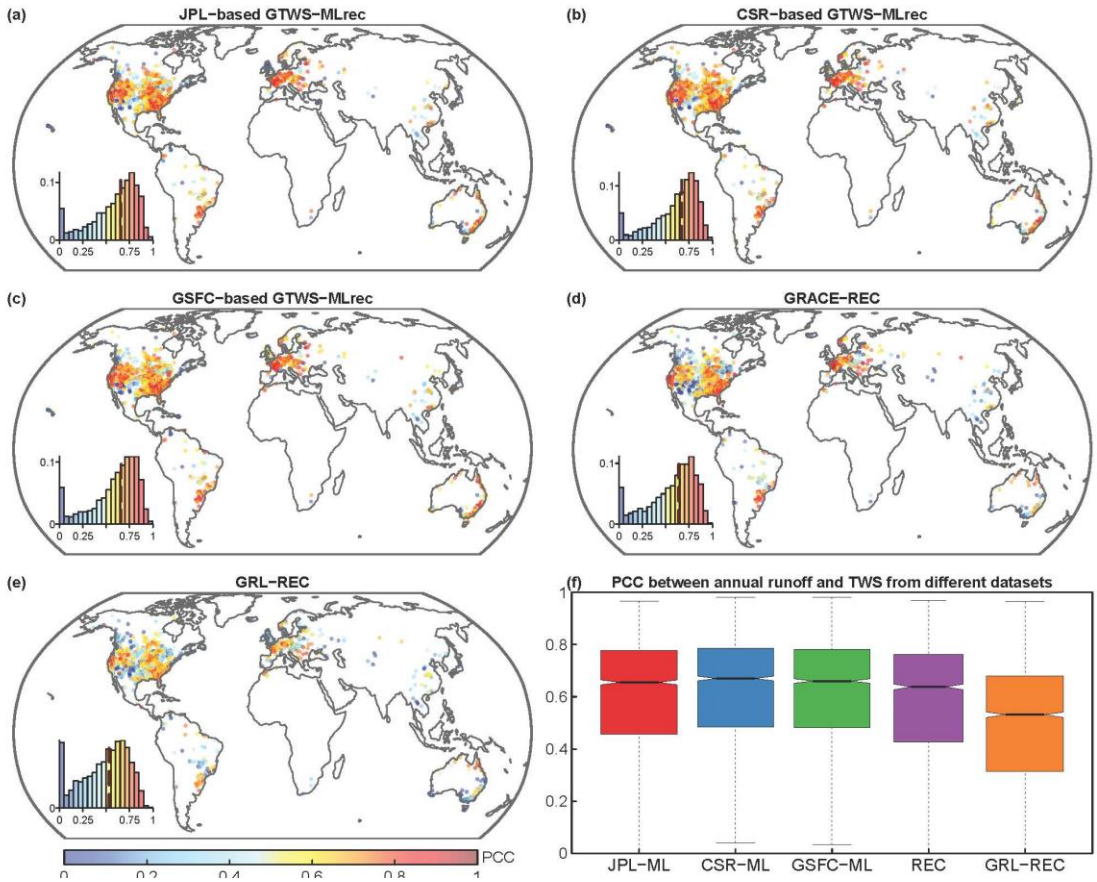

Figure 8. Correlation of annual streamflow and aggregated basin-scale TWS from the different reconstruction datasets during 1979-2022. **a-e**, Global distribution of PCC for the different datasets; **f**, Boxplot of the PCC for all stations globally; the REC denotes GRACE-REC. Insets in each figure show the histogram of these metrics, and the vertical red line shows the median value.

Reviewer #2

The article, titled "GTWS-MLrec: Global Terrestrial Water Storage Reconstruction by Machine Learning from 1940 to Present," presents a comprehensive and relatively high-resolution monthly time series of Terrestrial Water Storage (TWS) anomalies across the global land surface. This is accomplished through the application of a set of machine learning models, which incorporate climatic and hydrological variables, land use/land cover data, and vegetation indicators as covariates. The article is well written, and the methodology employed is well-defined. It aligns appropriately with the scope of the journal, and I am of the opinion that, with some minor revisions, it can be considered for publication.

Reply: We appreciate the reviewer's positive evaluation of our manuscript and the insightful comments on the improvement of the manuscript. All the concerns have been addressed in the revised manuscript; we hope the revised manuscript satisfies these concerns for publication!

I would appreciate it if the authors could provide further elaboration in the Methods section regarding the nearest-neighbor moving window approach, particularly mentioning the implications of choosing a 5x5 window size on the product's performance.

Reply: We have provided more details in the *Section 2.4 Machine learning-based TWS reconstruction method* as follows:

Before establishing the TWS reconstruction model at each pixel, a moving-window nearest-neighbour approach is employed to select the most important variables for each pixel and its immediate neighbours. The moving-window nearest-neighbour approach is a good method to improve the robustness of machine learning methods, and it can also improve the training dataset for calibrating the machine learning model by assimilating richer information from nearby points. To balance the size of data sample and model complexity, we use a moving-window size of as 5 × 5 for each pixel. We also tried a 3 × 3 moving-window size, and found it was slightly less robust than the 5 × 5 scheme.

Additionally, it would be valuable if the authors could explain the rationale behind opting for the second-best performing scheme to fill gaps in cases where the best-performing scheme may not be applicable.

Reply: We have provided more details in the *Section 2.4 Machine learning-based TWS reconstruction method* as follows:

In cases where variables might be missing for certain time periods, the best-performing scheme cannot be applied. To solve this issue, we make full use of the eight data schemes. For example, the best-performing scheme of the selected machine learning model is used to produce the TWS reconstructions, and then the second-best performing scheme for the same model is used to fill any missing gaps of the former one.

Furthermore, I suggest incorporating an analysis using the modified Kling-Gupta efficiency and its components in the evaluation process. This would help assess whether the models can adequately represent the correlation, bias, and variability ratio between simulated and observed data.

Reply: We have evaluated the performance by using many metrics, including Pearson's Correlation Coefficient (PCC), Nash-Sutcliffe efficiency coefficient (NSE); **Kling-Gupta Efficiency coefficient (KGE)**, Coefficient of Determination (R^2), Root Mean square error (RMSE, unit: mm), normalized Root Mean Square Error (nRMSE), Mean Absolute Percentage Error (MAPE), and the Percent bias (Pbias, unit: %). We think the above metrics already demonstrate the performance well in terms of correlation, bias, and variability ratio.

Readers would greatly benefit from a more comprehensive analysis of the performance of the various machine learning algorithms employed in the study. Additionally, it would be interesting to understand which variables had the most significant impact on improving their performance and to gain some insights into how and why these covariates influence the machine learning algorithms.

Reply: Thank you very much for this comment; we did compare the selected best-performing machine learning models in reconstructing the TWS series. We have also explored the most important variables in improving TWS performance. The LAI and relative humidity are the two key variables for improving the performance. But these contents are beyond our main objective, so we would like to keep our manuscript easier to read and will provide more systematical analysis in future work.

Just being curiuous here, but how the final resolution of the reconstructed products might influence their performance?

Reply: In some small-area catchments, a finer resolution of TWS can better represent the hydrological features.

In the evaluation based on water balance, there appears to be a correlation between catchment size and the Pearson Correlation Coefficient. Could this be related to the fact that the footprint of GRACE affects the water balance evaluation in relativelly small catchments?

Reply: Yes, we find that larger catchments might show higher PCC. We have provided a brief explanation in the *Section 4.3 Performance evaluation based on water balance over large river basins* as follows:

The larger catchments typically show higher PCC than the small catchments, which can be explained by the fact that the basin averaged TWS series of larger catchments have been smoothed by using more gridded data.

Minor Comments:

Line 23: Consider using "relatively high resolution" instead of "high resolution" for greater precision.

Reply: Thank you; we have revised accordingly.

Lines 262-264: There is repetition in this sentence. Please refer to the sentence between lines 250-251 for clarification.

Reply: Thank you; we have simplified the sentence in Section 3.1 Six GTWS-MLrec datasets as follows:

We also provide de-seasonalized and detrended TWS anomalies, which are independently reconstructed by using the de-seasonalized and detrended GRACE/GRACE-FO dataset and inputs.