Response to Reviewer No #1

We would like to thank the anonymous reviewer #1 for the detailed analysis of the manuscript and constructive comments. In this document we reflect on the comments made by the reviewer and how we propose to change/improve the manuscript in response to the issues raised. These comments are included for convenience (in blue font) as well as a detailed response to the comment and changes proposed to the manuscript.

In this study the authors use GRACE JPL mascon data to evaluate simulated total water storage (TWS) for 10 land surface (LSM) and global hydrological models (GHMs) over the Magdalena-Cauca basin (Colombia) and its sub-basins. They find different abilities of the different models to represent trends, seasonality and monthly time series, with model accuracy reducing from trends/seasonality to time series, from higher to lower resolution models and from larger to smaller basins. One of the models is declared the overall winner of the comparison.
I have the following comments:

Although this is an interesting and worthwhile exercise in itself, I am a bit hesitant about the novelty of this study. What exactly are the general conclusions we can draw from applying specific models to a specific basin? Global comparisons have been made before, as also testified by the references in the paper (Scanlon et al., 2016;2018; Schellekens et al., 2017). What does a regional study add to that? Does a study like this fit a general purpose hydrological journal like HESS, or does better fit a more applied journal that publishes well executed case studies? I leave it up to the editor, but if it is accepted, the authors should make clear what is novel about this work.

We thank for reviewer for noting that the study is of interest and worthwhile. We appreciate the point raised as to the novelty of the study, and agree that there have been several global comparisons (as referenced in the comment), as well as comparisons for individual basins (references provided in the paper). However, our study adds a novel contribution to these previous studies in two key ways. First, through developing a more detailed comparison at the basin level, which includes nested basins, allows a more comprehensive consideration of the importance of heterogeneous hydrological processes across the basin on the performance of these models against the benchmark provided by GRACE. This includes the poor representation of the dynamics of the wetlands in the lower basin, as well as the dynamics of TWS in the complex terrain of the upper basin. A second dimension that this study contributes is that the Magdalena-Cauca basin is unique amongst more detailed basin studies of GRACE as it is a tropical basin with a dominant monsoonal climate and pronounced ENSO influence (in parts of the basin), that also has a reasonably extensive and publicly available observed hydrometeorological dataset. This allows insight into model structures in LHM and GHM that would appear to provide a better representation of hydrological processes relevant to tropical basins to be obtained, and thus contribute to improving these also for basins in other tropical areas of the world that are not as well endowed with observational data.

To emphasize the contribution of the manuscript we propose the following changes:

Lines 9-10: a medium sized tropical basin with a well-developed gauging network when compared to other basins at similar latitudes.

Line 20-24: We conclude that GRACE provides a valuable dataset to benchmark global simulations of TWS change, in particular for those models with explicit representation of the internal dynamics of hydrological stocks, offering useful information for the continued models improvement in the representation of the hydrological dynamics in tropical basins around the globe.

Line 110-118: With the purpose to contribute to the understanding of the dynamic nature of TWS as well as to contribute to future LSM and GHM development and improvement, this study highlights the value of using water storage from GRACE, in addition to traditional water fluxes, as a benchmark in assessing global models in a tropical basin. The Magdalena-Cauca basin offers a special opportunity to compare global models for tropical basins, it has a dominant monsoon climate and a pronounced ENSO influence in some parts of the basin, and it also has a reasonably extensive and publicly available observed hydrometeorological data set. On the basis of Magdalena-Cauca tropical basin is unique amongst more detailed basin studies of GRACE, this assess allows insight into model structures in LHM and GHM, and contributes to improving these also for basins in other tropical areas of the world that are not as well endowed with observational data. Assessing models using these recently available data of an important state variable such as TWS, can contribute to a better understanding of the hydrological cycle processes, with the improvement in the modeling and forecasting of hydrological variables in tropical basins, thus being conductive to better tools for decision-making around water management and sustainability. The relatively large set of LSM and GHM models considered in this study are obtained through the open access global Water Resources Reanalysis dataset developed in the eartH2Observe (E2O) research project, a collaborative project funded under the European Union's Seventh Framework Programme (EU–FP7) (Schellekens et al., 2017).

Using GRACE that for sub-basins below 40000 km2 in size is very tricky, even if mascons are used. The inherent resolution of GRACE is too coarse for this. This means that the results for the smaller basins are questionable at best, and the differences between GRACE partly from the models and partly from the GRACE estimates. The question then is which part of the deviation comes from the models and which part from GRACE. The authors should either leave out the smaller basins or be very upfront about this limitation in the Introduction/Methods section already and not wait until the Discussion.

We agree with the reviewer that using GRACE for sub-basins below 40000 km2 is very tricky and indeed raise this in the discussion. In our opinion, leaving the smaller basins out of the analysis would quite substantially change the contribution of the manuscript. As such we follow the reviewer's suggestion to be quite up-front about this.

We propose to add a sentence in the introduction in lines 83-88:

Line 83-88:
Although GRACE has important limitations due to its resolution (Chen et al., 2016), data from GRACE do provide a uniquely independent estimate of the distributed TWS in a river basin as water balance estimation based on observed data and models require gauging data (which are often deficient or insufficient) or data from reanalysis models, which are not direct observations. Advances in GRACE processing from traditional spherical harmonics to more recent mass concentration (mascon) solutions have increased the signal-to-noise ratio and reduced uncertainties (Scanlon et al., 2016), though the interpretation of results for basins smaller than on the order of ~40,000 km2 remains difficult due to the inherent coarse resolution of GRACE data (Scanlon et, al. 2016; Vishwakarma et al., 2018).

Line 33-35: The argument that knowing TWS leads to better forecasts is often used. Please provide us with examples from the literature where it is shown that significantly better streamflow forecasts are obtained when GRACE TWS is ingested into the model?

Our comment here is that a better representation of the basin initial state contributes to improved (streamflow) forecasts, in particular in basins where there is significant internal storage and persistence of initial states. Examples that have shown the specific contribution of GRACE TWS in improving streamflow forecasts include a recently published study by Liu at el, 2021 (https://doi.org/10.1080/02626667.2021.1998510), as well as Getirana et al 2020 (https://doi.org/10.1175/JHM-D-19-0096.1). These references will be included.

Line 35-42: I have to say that this argumentation is a bit silly. Before GRACE, nobody cared about the validating TWS of hydrological models at all! The reason is that it could not be observed. Before GRACE, only partial state variables, such as groundwater, river and lake levels, soil moisture and SWE were independently evaluated using in-situ and remotely sensed data. Only after GRACE, TWS anomalies could be validated and were therefore computed from models.

We appreciate that before the availability of GRACE the interest in TWS as a variable may not have been apparent as it could not be observed. However, the point we had intended to raise is the importance of monitoring or estimation of change in water stocks as represented by TWS to water resources assessment, and that given the difficulty of independent integrated observation, water resources assessments are commonly developed using models and water balances.

Despite the acknowledged importance of this variable, prior to the availability of data from GRACE integrated observations of water stocks at the basin scale, as represented by TWS, were unavailable, with only partial state variables such as groundwater levels, soil moisture, river and lake levels, and snow water equivalent available from direct in-situ and remotely sensed observations. Given the heterogeneity of the hydrology of river basins, comprehensive observation of these is, however, very difficult due to insufficient in-situ observations of these partial variables, further confounded by the global decline in gauging networks (Hassan and Jin, 2016). Estimation of TWS and its change at the basin scale is therefore commonly done through water balances and the use of models. Given the difficulty to measure TWS (Tang et al., 2010) these. Many traditional analyses assume that at longer timescales and over large regions, change in TWS can be approximated as zero. This implies that in water balance studies it is common to ignore the long-term trends of TWS (Reager and Famiglietti, 2013).

Figure 11 shows that WaterGap and Lisflood both show relatively poor performance in reproducing TWS anomalies. What is striking is that these models both have been subject to some sort of calibration to streamflow data (see the paper by Beck et al., 2017 where they perform very well in streamflow reproduction). Could it be that calibrating GHMs to streamflow only (without constraining internal states and fluxes by other information) has led to correcting errors in streamflow by accruing errors elsewhere in the model?

We thank the reviewer for underlining the poor performance of these two models, in particular, for the R1 dataset. The performance of WaterGap in the R2 dataset is also quite clear, and is discussed extensively in the discussion (note that LISFLOOD was not available using in the R2 dataset). We agree that the introduction of calibration of models against observed discharges may improve model results at the basin outlets, which may be detrimental to representation of internal states and fluxes. This may be particularly so where there are biases in the forcing data over the complex topography of the basin (see also response to Reviewer #2). The paper of Beck will be added to improve the discussion.

References

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881–2903, 2017.

Scanlon, B., Zhang, Z., Save, H., Sun, A., Schmied, H., van 630 Beek, L., Wiese, D., Wada, Y., Long, D., Reedy, R. C., et al.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, Proceedings of the National Academy of Sciences, 115, E1080–E1089, 2018.

Scanlon, B., Zhang, Z., Rateb, A., Sun, A.,Wiese, D., Save, H., Beaudoing, H., Lo, M., Müller-Schmied, H., Döll, P., et al.: Tracking seasonal fluctuations in land water storage using global models and GRACE satellites, Geophysical Research Letters, 46, 5254–5264, 2019.

Schellekens, J., Dutra, E., la Torre, A. M.-d., Balsamo, G., van Dijk, A., Weiland, F. S., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., et al.: A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset, Earth System Science Data, 9, 389–413, 2017.