Response to Reviewer No #2

We would like to thank the anonymous reviewer #2 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.

This manuscript highlights the use of an external source of data (JPL GRACE TWS monthly anomalies) to benchmark 10 different global hydrological and land surface models using results of the earth2observe project, in a well-instrumented tropical basin in Colombia, the Magdalena-Cauca (MC) macrobasin as the area of study. Findings identify characteristics and limitations of the models and are a key input for contributing to identifying new developments and improvements of these types of models.

The article is well written, organized, and discusses nicely the main findings. The objectives the paper sets out to are of interest, and there is scientific merit for publishing it. Below are some specific comments to the authors:

We thank the reviewer for the positive comments on the scientific merit and structure of the article.

In the abstract (line 11) and the methodology, analysis and long-term tendencies in terrestrial water storage (TWS) are based on JPL GRACE data from 2002-2014. What are the limitations of these estimations taking into consideration that the period is short (only 13 years), that the MC has a large inter-annual climate variability associated with the ENSO and other phenomena, and that the base period used to calculate the anomalies is also short (2004-2009)?

We thank the reviewer for the comment, and agree that the length of the data is not that long, but by long term tendencies we actually mean variability at the ENSO time scales (i.e. multi-annual to decadal time scales) and not longer than that (e.g. climate time scales). According to the article Bolaños et al, 2020, in the long-term analysis, the effect of climate change on TWS is not conclusive due to the short time series. Therefore, the observed trends may be mainly due to the climatic variability present in the region.

This study focuses more specifically on long-term variability. In the study, the long-term series is obtained through Seasonal Trend decomposition by Loess, (STL) proposed by Cleveland et al. (1990) to estimate the relative magnitudes of water storage variance of different time series components. This is done in order to compare the seasonality and long-term trend of the models with respect to GRACE, and to observe how they capture the climatic variability in the tropical region. With the new data collected by GRACE Follow-On, it will be possible in the future to have a longer TWS series that allows better long-term analysis.

Although it is not completely clear in the manuscript, because it is not explicitly mentioned in the Data and Methods section, it seems (see line 103, line 221) that TWS is calculated from the models´ results and the JPL GRACE data at the macrobasin and subbasins scale using the average of the values for all the cells in the corresponding domain and time step. If this is true, this approach could have some limitations that the authors should address within the discussion and conclusions. And if not, an explanation of the methodology used and its limitations should be included in the manuscript.

In fact, the TWS anomaly data was averaged within the time step and domain.
So, we add the follow sentence in Line 196 to clarify the method: For both TWS calculated from models and JPL GRACE data, the values of all cells were averaged corresponding to each time step to construct the time series for each database and each watershed. This implies some limitations such as sensitivity to extreme values, and the averaging of biases, therefore, in this study we assume that the variables are normally distributed.

In lines 170 and 418 it is important to consider that from WRR1 to WRR2 some models also have some type of calibration, not necessarily in the MC basin.

We thank the reviewer for the comment, and agree that it is important to mention that indeed some of the models in both WRR1 and WRR2 are calibrated. Models that have been calibrated include LISFLOOD, WaterGAP, HBV-SIMREG and SWBM (see also comments by reviewer #1 and Beck et al., 2017b).

We add the following sentence in line 170:
Selected models in WRR2 and WRR2 have been calibrated against streamflow data, including LISLFOOD, WaterGAP, HBV-SIMREG and SWBM (Beck et al., 2017b).

We also include an indication of the models that have been calibrated in Table 1.

The sentence in line 418 has been amended:
In this study, the models have the same input/forcing for each WRR and the majority of models have not been calibrated, while for those models that have been calibrated (Table 1) these calibrations were not specific to the MC basin (Beck et al., 2017b). This implies that differences must largely be due to model structure (e.g. representation of water storage compartments) and parameterization (e.g. capacity of compartments) (Dutra et al., 2017).

The legend used for the different models and modelling phases (WRR1 and WRR2) is consistent throughout the document. However, the first time the legend is introduced is in Table 2. Perhaps an explanation of the legend in this Table would facilitate the analysis right from the beginning of the paper.

We have included the legend by which models are identified in Table 2 (e.g. R2eq1) and have also amended the title of the table to clarify.

Table 2. Components used in TWS change estimation for each model. The last four columns included the symbols and legends used to identify model resolution and equation used to derive simulated TWS change. These are used throughout the manuscript.

Equation 3 proposes a way to decompose the time series of TWS into seasonality, long term, and residuals. For the first two components, a detailed analysis is conducted. However, for the residuals, it is not the case. The analysis of the residuals would be a nice way to complement the findings of the study.

In line 335 perhaps the analysis of the residuals quite nicely complements the results.

Equation 3 proposes a way to decompose the time series into seasonality, long-term, and residuals through the Seasonal Trend Loess (STL) decomposition. However, the main objective of the study was to compare the seasonality and long-term variability of the models with GRACE.

Much of the time series variability that we can observe is included in seasonality. Furthermore, an analysis of residual correlations was done, and the results were very poor, so we decided not to include it in the study.

In Figure 2 they appear 7 different GHM including SWBM_Exp 1 (in addition to SWBM). This experiment with this model is not described either in Table 2 or in the text. For consistency in the document, where 10 models are analyzed, this experiment should be dropped from the analysis.

In Figure 11 it also appears the SWBM_Exp1 model, which either should be described in the

manuscript considering 11 instead of 10 models or dropped from the analysis.

Thank you for this observation. We accept the suggestion.

In previous studies that have used discharge to investigate the performance of the models in the earth2observe project in the MC basin it has been shown that LISFLOOD obtains the lower results as it is also confirmed in this study (line 239). Reasons for the low performance of this model in the MC are not discussed in the document and would be helpful to include.

Thank you for this observation. In fact, LISFLOOD obtains lower results, and one reason could be that LISFLOOD is reported to underestimate quick flow response, which could lead to less pronounced seasonality (as shown in Figure 9). PCR-GLOBWB is also reported to have this issue (see Beck, 2017b). On the other hand, this poor behavior could also be due to the calibration.
The discussion of this will be included in the document as de reviewer suggests.

In several parts of the article a threshold of 60,000 km2 has been proposed as the basin size limit for the use of GRACE data to validate the models. In this sense would be the Cauca (C) basin an exception? How do the different climatological regimes in the C and Upper Magdalena (UM) basins influence the results? It is evident that for the small basins including UM, Upper Magdalena Paez (UMP), and Saldaña (S) results are poor and this is the reason for choosing the size limit proposed. However, right from the start results in the UM are poor, so for other subbasins in this area, it would be expected that results are also poor. What would happen if instead of considering subbasins in the UM you choose subbasins in the C (additional to the Upper Cauca (UC), where the size is small and surely below the limit), where results are much better?

In the study, only five subbasins have drainage areas close to or below 60,000 km2. Considering the climatological and physical complexity of the MC macrobasin, in my opinion, there is not enough information to establish the threshold proposed as a basin size limit for evaluating model performance against GRACE data.

In this study, we are not suggesting a threshold as such, but we do observe a marked difference for the basins above/below this basin size. In the bibliography, Vishwakarma et al., 2018 propose a basin size limit of ~63,000 km2, which is close and consistent with what was found in this study. On the other hand, although in UM, UMP, and S the behavior is poor (such as being expected since UM subbasin is poor), we consider them in the study because they comprise the upper area of the basin, which has complex topographic characteristics and the presence of moors, which models fail to represent adequately.

Following the previous comments, for the UM and C basins, with approximately the same size, there is quite a contrast in the results. For the first subbasin, results are way lower than for the second one. Similar results in the UM that the ones presented in the study have been obtained with several different models, not only global but also regional and local. In this sense, any model structure seems to perform poorly in the UM. Problems in the precipitation forcing used for this basin could be part of the reason? Recent studies (unpublished) have shown that in some basins of the UM, including the S, the monthly precipitation and discharge average patterns do not match. Rainfall is mainly bimodal, as captured by the models´ forcings in this study, but streamflow is mainly unimodal. This could be associated with anthropogenic interventions, clearly discussed in the manuscript, but also with climatological forcing limitations that need to be addressed in the paper.

In line 448 besides the reasons for the poor performance of the models in the UM, perhaps influence from the Orinoco and Amazon macrobasins, may also play a role in the results. Some consideration about this is also recommended to be included in the discussion.

Thank you for this observation. We included this in the document as reviewer suggests.

We add the following sentence between line 441 and 442.

Line 441: … but is also influenced by other atmospheric circulation mechanisms such as meso-scale convective Systems, soil-atmosphere interaction processes, and local circulation patterns (Poveda, 2004). Furthermore, the interplay between the Orinoco and Amazon basins plays an important role in the moisture availability for UM precipitation, which would add complexity to the hydro climatological characteristics in this region, and a huge challenge for models and their calibration.
It is important to highlight in the UM basin, the presence of high altitude montane wetlands (Paramos)…

In line 274 it should be Figure 4c instead of Figure 5c

This has been changed as suggested.

Figures 5 and 6 (line 282, line 309) in my opinion could be included in the supplementary material, as they are not key for supporting the main findings described in the article. Instead, the analysis of the residuals perhaps could better support the analyses and discussion.

We agree with the reviewer and have moved these to the supplementary material as well as the references to these figures.

In line 283 it should be In these figures… instead of In this Figure ….

This has been changed as suggested.

For Figure 4 there is enough space in the graph to include the accompanying legend to facilitate the interpretation of the results.

We have added the legend to Figure 4.

Sentence in line 321 is not clear.

The sentence has been amended as follows:
… which means that these dry out too much in the drier DJF period and cannot represent the increased storage in the wet period in SON in La Mojana area, which can be observed by GRACE.

Results for the WATERGAP3 model in the Limpopo River Basin have shown the good performance of this model (line 356). Results in the MC and some of its subbasins have also shown good results for this model when discharge observations are used. How to interpret that when using GRACE data as a complementary source of validation, results for this model deteriorate so much?

We infer that this could be due to the calibration and representation of internal model states. The model could adequately represent the discharge but the internal hydrological processes could not be right to underestimate/overestimate some fluxes, which means that the discharges are reasonable, but that this is to the detriment of internal model states. Since TWS is a state variable, it gives an approximation of the internal hydrological processes that result in discharges in the basins. Therefore, including TWS in the models can improve the representation of flows without detriment to their internal states.

Instrumentation in the UM, especially in the higher altitudes could in my opinion help to separate the influence of the anthropogenic interventions from the limitations in precipitation forcing and how they impact the streamflow patterns observed for this part of the MC catchment.

We agree with the reviewer y we thank you for this comment. In the document, we discuss that UM poor performance could be due to different reasons or the sum of all limitations in hydroclimatologic calibration, complex topography it this region, and influence of the anthropogenic interventions in the basin. On the other hand, the Cauca basin could be more strongly influenced by climate teleconnections which could contribute to its better performance in contrast with UM.

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, 2017b.

Bolaños, S., Salazar, J. F., Betancur, T., and Werner, M.: GRACE reveals depletion of water storage in northwestern South America between ENSO extremes, Journal of Hydrology, p. 125687, 2020.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: a seasonal-trend decomposition, Journal of official statistics, 6, 3–73, 1990.

Vishwakarma, B. D., Devaraju, B., and Sneeuw, N.: What is the spatial resolution of GRACE satellite products for hydrology?, Remote Sensing, 10, 852, 2018.