Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-323-RC2, 2021

Comment on hess-2021-323

Anonymous Referee #2

Referee comment on "Benchmarking global hydrological and land surface models against GRACE in a medium-sized tropical basin" by Silvana Bolaños Chavarría et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-323-RC2, 2021

Review of Benchmarking global hydrological and land surface models against GRACE in a medium-size tropical basin, Bolaños et al.

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:

- 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)?
- 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 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.

- 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.
- 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 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 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.
- 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.
- 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 274 it should be Figure 4c instead of Figure 5c
- 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.
- In line 283 it should be In these figures… instead of In this Figure ….
- For Figure 4 there is enough space in the graph to include the accompanying legend to facilitate the interpretation of the results.
- Sentence in line 321 is not clear.
- In line 335 perhaps the analysis of the residuals quite nicely complements the results.
- 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?

- 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.
- 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.
- 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.