Text in bold: Reviewer comments; *Text in italic: Authors response*

Anonymous Referee #1

This m/s presents a continental hydrological model for South America. It is forced by global rainfall and climate data and calibrated to streamflow records for a large number of stations. The agreement with recorded streamflow is presented, as well as that with satellite-derived evaporation (ET) and total water storage (TWS). The agreement with observed streamflow is better than that of an ensemble of 3 global models driven by the same precipitation estimates.

Overall assessment: This appears an overall competent and sound study, but I am missing some truly new scientific insights. The abstract suggests the main insights are (1) calibrating rainfall-runoff parameters is necessary to simulate discharge appropriately; and (2) implementing hydrodynamic routing is also important. I don't think either of those is really very novel. I do not think there was ever any doubt that parameter calibration against streamflow records was going to improve the agreement with those same records (noting that the "appropriately" used in the abstract is obviously a subjective term, or at least one that would have to be purpose-specific). The second conclusion also hardly seems surprising and has been shown in previous studies, specifically for the Amazon basin. Indeed, the authors provide several literature references that offered those very conclusions.

We are grateful to the reviewer for reading our manuscript and pointing out relevant questions that need clarification.

To the best of our knowledge, this is the first study that provides a comprehensive cross-scale comparison between regional/continental x global models, supported by spatially distributed, daily discharge data. As stated by the reviewer, there is no doubt that calibration against discharge records leads to improvements in this same variable. Intercomparison studies indicate that models with some degree of parameterization generally perform better on average (e.g., Zhang et al. al., 2016, Beck et al., 2017). On the other hand, results from this paper put light on what extent both calibration and improvement in model physics (routing) are expected to improve estimates of daily river discharges, focusing on different aspects (overall agreement, high and low flows, timing, bias) and regions of South America. To our knowledge, no other study provides such discussion.

Although these results do not necessarily show the performance limit that can be achieved by current global forcing data, we understand that these results are of interest to the modeling community that has been seeking for locally relevant hydrological estimates, especially in under development regions. Recent studies call for cooperation between scales (e.g., Archfield et al., 2015) and this work is going exactly on that direction. In summary, continental-scale modeling shows that it is possible to get better discharge estimates by using global data and methods that are currently available, as well as knowledge and methods developed for the studied region. Therefore, we believe

that our manuscript brings new information that is relevant in the context of regional, continental and global-scale hydrological modeling.

We agree with the reviewer that the term "appropriately" in the abstract is a subjective term. We also understand that some modifications in text are necessary to clarify the aforementioned statements.

On the positive side, I do think this is an interesting study that has the potential to be a valuable contribution. I thought some of the most interesting contributions from this study were:

1) **There is a much larger set of streamflow gauging stations in existence in South America than is represented in global databases and typically used to calibrate global models;**

Thanks for this comment. Yes, much of the available data are hardly accessed by modeling studies with global coverage. As shown in section 3.2.1, we have only selected gauge stations with drainage areas above 10000 km², that is, much data is still available to be used in the future for more detailed evaluations. It is worth mentioning that a lot of time is needed for data acquisition, quality check (time series, gauge location, drainage area) and standardization considering different institutions, and these efforts can put constraints on data usage. But we understand that handling local information is important since global databases sometimes suffer from "disinformative data" (Kauffeldt et al., 2013).

2) The use of a large number of altimetry-derived water level records is interesting;

Thanks for this comment. Indeed, the use of satellite altimetry is interesting to evaluate the performance of the routing component (in addition to ET and TWS that are used to evaluate the water balance), especially in regions where data are scarce.

3) The authors provide some interesting commentary on the hydrological conditions that likely explain consistently poor performance by global models in some of the basins in South America.

Thank you for taking a positive look at this discussion in the manuscript. This point, for example, comes right into the idea of reducing the gap between global and regional modeling. Studies focused on the global scale are usually not concerned with specific processes occurring in specific regions, but there are key studies - many of them on a regional scale - that provide further insights about modeling results and can partially explain the performance obtained. Therefore, we believe that studies at the continental scale can serve as a link for communicating relevant findings to a broader audience, helping to improve the general understanding.

I was somewhat surprised that the majority of forcing and spatial parameterisation approaches used for the "regional" MGB model were, in fact, the same as used for global models. Furthermore, Fig. 11 appears to suggest that the inclusion of hydrodynamic routing was a minor factor in explaining the generally better performance. Therefore, it would seem that the larger number of streamflow gauges and their good use in a more intensive calibration were the real reasons for better performance. That in its own right is useful, as it sets a benchmark that global models should be able to achieve with appropriate parameter calibration.

What is unclear, however, is whether that would go at the detriment of the agreement with other observations of the water cycle, such as ET and TWS. It is common that a heavy emphasis on streamflow calibration leads to deterioration in other terms. Therefore, I was surprised that the authors did not include the global models in their comparison against ET and TWS, to assess whether those were simulated better or worse. (The altimetry water levels are less relevant in such a comparison, as one would assume that better discharge simulation also produces better water level simulation. Nonetheless, a comparison with the global models might still have been of interest.)

Thanks for the comment. Yes, many of the techniques and databases used to prepare the continental model are similar to those used by global models. Such strategy indicates that the latter can achieve similar results even without a significant increase in the number of computational elements. As pointed out by the reviewer, parameterization has a greater impact on model results in comparison to improved routing, although efforts in calibration may be not effective in regions such as the Paraguay basin. Another problem is when a simplified routing structure is applied together with calibration focused in gauge stations with large drainage areas, especially in basins affected by floodplains (e.g., Amazon at Obidos). This can cause problems on flow timing at upstream regions as discussed on the manuscript.

Regarding the comparison between the continental x global models using variables such as ET and TWS the reviewer makes an interesting suggestion, but we understand that this would not fit into the context of the present study. We have focused on river discharge because this variable is widely used for water resources planning and practical applications (e.g., hydrological forecasting, reservoir operation). It would be difficult to demonstrate if a given model has better accuracy than another with respect to ET since the reference (CDR) has large uncertainties. In addition, to our knowledge, TWS for WRR-2 is not directly available for each model, but rather for an ensemble mean generated with the inclusion of several other global models (i.e., not only those used in this study). Yet, even if there is some interest in comparing water level anomalies, this would not be possible because water level is not an output of WRR-2.

We agree that calibration against discharge records could lead to decreased performance in other hydrological variables such as ET and TWS. This is the reason

why we performed several evaluations of both variables seeking to document model errors and to understand potential sources of uncertainty.

In summary, the present m/s mainly seems to assert the common "our model is better than theirs", which is not very insightful as it appears almost entirely due to calibration. There are however some good opportunities to make this a more valuable (and cited!) contribution:

Thanks for this comment. We will review the manuscript to make sure that analyses and discussions are constructive and not biased.

Model intercomparison studies sometimes make direct comparisons of performance (e.g., Xia et al., 2012, Zhang et al., 2016, Beck et al., 2016, Beck et al. 2017), and we understand that this is important because identifying model shortcomings is essential for future improvements to be made. In this context, there is no best model (first because it depends on the objectives) but rather issues of scale and the best use of available information. For instance, it is expected that models with a regional domain (even if forced with global data) will have, on average, better performance when compared with a continental model. Of course, it would be of high interest to understand which areas/rivers, flow conditions (high and low flows, bias, timing...) and for what reasons this happens, which definitely represents a reduction of the gap between different scales. The reviewer seems to agree with this idea when he/she states: "The authors provide some interesting commentary on the hydrological conditions that are likely to explain consistently poor performance by global models in some of the basins in South America."

1) Propose these model simulations, along with the station and altimetry records, as a benchmark for global models by making them directly available online to the global modelling community. To make the MGB model acceptable as a benchmark for an allround hydrological model, you should demonstrate whether the global models are also less effective in simulating ET and TWS. This would provide insight into whether only the streamflow simulations can be considered benchmark, or the other water cycle components as well.

Thanks for the suggestion. Indeed, model outputs will be available for public access in a specific website or by request to the authors. This will be indicated in the item "data availability" in the revised version of the manuscript. In addition, we can make available a list of the gauge stations in the supplementary material (name, coordinates, drainage area, respective institutions), since these data can be downloaded according to section 3.2.1. We can also recommend the development of a dataset to facilitate both model validation and intercomparison in South America, as suggested by Reviewer #2.

With respect to ET and TWS, we have previously indicated the reasons why it would not be feasible to make such a comparison at this moment, apart from the fact that the current study is already long.

Provide more discussion and emphasis on the understanding of the hydrological conditions of some of the "problem" basins, so that they might become a valuable "stress test" of hydrological model performance.

Thanks for the suggestion. We will add discussion in the revised version of the manuscript. This will certainly bring contributions for proposing ways of improving future modeling applications (both regional and global).

References

Archfield, S.; Clark, M.; Arheimer, B; Hay, L. E.; McMillan, H.; Kiang, J. E.; Seibert, J; Hakala, K; Bock, A.; Wagener, T.; Farmer, W. H.; Andréassian, V.; Attinger, S.; Viglione, A.; Knight, R.; Markstrom, S.; Over, T. Accelerating advances in continental domain hydrologic modeling. Water Resources Research, 51(12), 10078-10091, 2015.

Beck, H. E., van Dijk, A. I. J. M.,de Roo, A., Miralles, D. G., McVicar, T. R.; Schellekens, J., Bruijnzeel, L. A. Global-scale regionalization of hydrologic model parameters. Water Resources Research, 52, 3599-3622, 2065

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, Hydrology and Earth System Sciences, 21, 2881-2903, 10.5194/hess-21-2881-2017, 2017.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrology and Earth System Sciences, 17, 2845-2857, 2013.

Xia, Y., Mitchell, K.; Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., Lohmann, D. Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. Journal of Geophysical Research, 117, D03110, doi:10.1029/2011JD016051, 2012.

Zhang, Y., Zheng, H., Chiew, F. H. S., Arancibia, J. P., and Zhou, X.: Evaluating Regional and Global Hydrological Models against Streamflow and Evapotranspiration Measurements, Journal of Hydrometeorology, 17, 995-1010, 10.1175/jhm-d-15-0107.1, 2016.