Responses to reviewer comments

We thank the reviewers for spending the time for another review of our manuscript. Like last time, we will repeat the reviewer's comments (italic font) and response directly below (standard font).

RC1

The authors have revised the manuscript according to the reviewers' comments. They have also updated and extended text and figures to further improve the manuscript. I think the manuscript has developed a lot and I find most of my comments well addressed.

Thank you.

However, I still don't agree with the way the authors evaluate whether water storages have reached equilibrium in their model. In dry regions, where inflows, outflows, and equilibrium storage are small, a small residual trend in storage change does not necessarily indicate that storages have reached equilibrium. An example for a more suitable metric would be the ratio of residual storage change to the sum of the storage change and the storage outflow (which should equal the inflow). This provides a direct measure for how much of the inflow is still used to fill up the storage and, by that, by how much outflow is still affected. I understand that for the current study, it is not very critical to have all water storages in equilibrium. But since the authors have dedicated a relatively large portion of the paper on this issue, it would seem important to address it appropriately. I would like to see the shortcomings of assessing equilibrium based on absolute residual storage changes at least briefly discussed. Perhaps the authors could complement that by an estimate of how many grid cells are in equilibrium using the metric above, for example.

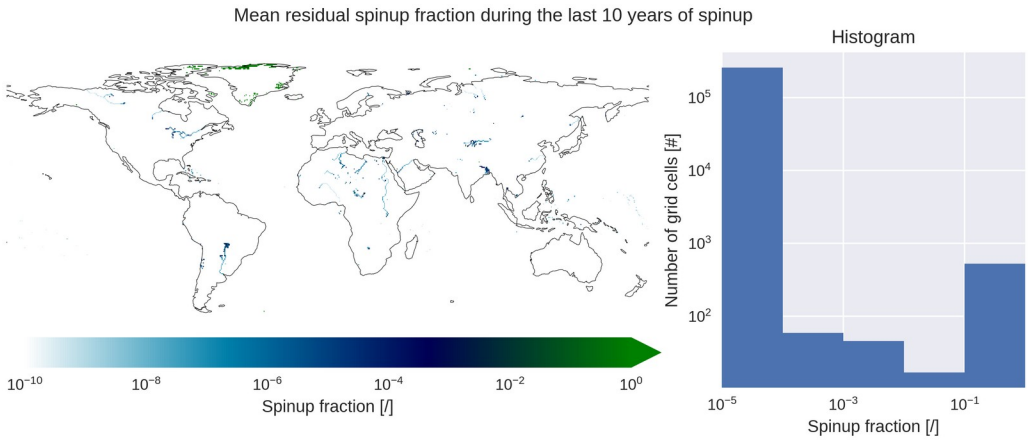
We assume that part of our disagreement on this point might be due to a different perception of the goal of this spin-up evaluation. Our reasoning is not to provide the model with a perfect initial state. Actually, this would not be the best initialization anyway, as many regions (e.g. desiccating lakes) are not in an equilibrium state in reality. Thus, the initial state we strive for is one that the model can use without:

a) experiencing any kind of initialization shock due to large changes in storages during the first steps of the simulation which might interfere with the results afterwards and

b) experiencing any residual trends which might then be wrongly mistaken for real signals.

For this reason, the state of very dry grid cells do not matter as much for our simulations (as the reviewer already acknowledged). Moreover, for the same reason, we indeed think that trends provide a good measure of the suitability of our initial state because their size and distribution tells us, whether any residual signals might be expected in our production simulation.

Nonetheless, we very much thank the reviewer for proposing this interesting alternative metric. Applying it for the last 10 years of our spin up simulation confirms our spin-up evaluation using trends. With the exception of the mentioned glacier cells where the spin-up fraction is still above 10%, there are less than 100 cells that show a residual storage change larger than 0.1% of the annual sum of outflow and storage change.



Furthermore, these cells are not located in dry areas, but rather along main river channels. Exactly the same signal pattern occurs in our trend analysis. However, they correspond to trends <= 0.1 kg m-2 a-1 and, thus, do not show up in Fig. 5 of the manuscript as we consider such trends to be negligible for our simulation and chose our color map accordingly.

Although we already mentioned in our manuscript that the spin-up target is the production of a suitable field for model initialization and not a real equilibrium state, we further modified the text and replaced all mentions of “equilibrium” with “stable state” to better reflect our intention with this analysis.

One other point I didn’t catch in my first review is the interpretation of model performance based on normalized Nash-Sutcliffe Efficiency (NNSE). The NNSE range of what is considered sufficient performance in Moriasi et al. (2007) refers to daily discharge time series. Applying these thresholds to NNSE calculated for monthly discharge climatologies is inappropriate and falsely implies a performance similar to calibrated watershed models. I think a clarification of the how NNSE values of monthly discharge climatologies are to be interpreted is needed here.

After carefully re-checking the study of Moriasi et al (2007), we do have to disagree on this point. Table 4 (page 891 in Moriasi et al. 2007), which is the source of our values, is even explicitly named: “General performance ratings for recommended statistics for a monthly time step.” On the same page, it is mentioned that “The model evaluation guidelines presented in the previous section apply to the typical case of continuous, long-term simulation for a monthly time step.” For this reason, we don’t think that we falsely imply a very good performance. Especially, as we don’t claim it globally, but just for a minority of catchments. Please note, that while working on your remarks, we actually found a bug in our NSE calculation, however, the 20% best-performing catchments were hardly affected by it.

Anyway, as there is a general trend to prefer the Kling-Gupta efficiency (KGE) over NSE (e.g. Knoben et al, 2019), we meanwhile changed our analysis setup to use the KGE instead and adapted our analysis accordingly. Thus, we updated those parts of our analysis that were based on the NSE (Sec. 4.3 and Sec 4.4), with the corresponding figures 8, 12 and 13. We also slightly modified the selection of river basins to discuss the differences between HydroPy and MPI-HM. Contrary to the NSE, there are no specific categories defined for the KGE but generally positive values are considered to indicate model skill (Knoben et al, 2019). For this reason, we removed the sentence comparing HydroPy to calibrated

catchment models. Note, that the switch from NSE to KGE does not change any of the conclusions of this study.