We thank the reviewer for his/her interest in our work and for the thoughtful and detailed feedback provided.

Comment:

This manuscript reports on a comprehensive calibration and validation experiment of a hydrological model at large spatial scales. The value of this manuscript is less on learning on a particular model, on the hydrology of a river basin, or on how a suitable and well performing hydrological model should look like for this particular region, but its value is much more on presenting a well-defined procedure of a step-wise multi-variable and multi-criterial calibration scheme towards improving model structure. The study illustrates the use of even coarse and uncertain remote sensing data, including the spatial patterns of state and flux variables (here water storage variations and evapotranspiration) in the calibration and model adjustment approach. In this respect, it provides a valuable example and guideline for other studies in future. I therefore recommend its publication in HESS after considering some comments as listed below

Reply:

We highly appreciate this positive assessment of our manuscript. We will in the following address all specific comments in detail.

Comment:

A thematic/scientific drawback of the study is that a significant improvement of the spatial patterns of simulated ET and storage anomalies could not be achieved within the set of model modifications tested here, in spite of some increase in the performance criterion. In particular, the pattern of areas with high ET in the remote sensing product could not be reproduced by the model. The authors argue that missing lateral sub-surface flow between modelling units could be a reason for this. Can a related modification of the model structure additionally be tested? A more convincing outcome in this direction could also be of benefit for the paper as a whole in demonstrating the value of the multi-criterial calibration approach on spatial patterns.

Reply:

This is indeed a very important point raised by the reviewer. It is true that even after rather extensive model testing and adaption, the representation of spatial pattern did only slightly improve. We also think that it is not implausible to assume that lateral exchange between the model units may explain some of these model deficiencies. Adding such lateral exchange to the model is of course possible, in principle. However, it is not a trivial thing to do in a meaningful way. We believe that this is in itself is a major research topic which warrants several in-depth research papers on its own and which cannot be done as a mere additional hypothesis in this analysis. The underlying reason for this is partly implicit in the nature of the model type used here and partly in the data that are available with current observation technology. Lateral exchange fluxes are (as any fluxes) driven by the interplay between continuous gradients and resistances. Conceptual type models are based on simplified expressions that mimic gradients *within* a model domain only. If such a model is implemented in a spatially distributed way, the individual model domains are the model grid cells. Gradients are thus only defined *within* these grid cells, but not across them. Thus, the head difference between adjacent model grid cells is undefined. In the absence of such gradients it therefore remains unknown between which grid cells such lateral exchange fluxes occur, into which direction and at which rates. As a consequence, these fluxes can only be expressed on basis of free calibration parameters. Depending on the degree of spatial discretization at the very least 4 additional free calibration parameters (Model E) would here be needed to represent exchange fluxes between the model grid cells. This does not yet include potential exchange fluxes of each grid cell with adjacent grid cells outside the Luangwa basin. In comparison, the fully distributed Model F would even require at least 30 additional calibration parameters. In the model calibration process, these additional parameters and the associated increase in the degree-of-freedom of the model, will very likely lead to improved model performances. This may even extend to the model validation period. Yet, the inclusion of such processes will not be warranted by the available data, as we will have no means of testing whether the additional calibration parameters and the associated exchange fluxes are physically plausible. We may end up with a model that features nice performance metrics for calibration and potentially also for validation, but in which water may flow against real-world elevation and/or pressure gradients or, to express it in a pointed way, water may flow uphill. These unspecified boundary fluxes across grid cells are at the core of the closure problem (Beven, 2006) and touch on the limits of what can be done in hydrology with our current observational technology and the available data. We will discuss this in more detail in the revised manuscript.

Comment:

In this respect, the authors discuss the dominance of the discharge performance criterion within the overall performance measure that was used for calibrating against all variables and criteria. Has the ability of the model to reproduce that spatial ET patterns been tested with varying weights among the different criteria in the overall measure, or for single-criterion calibration the ET patterns only? The (in)ability of the model to represent this feature and the trade-offs relative to other criteria could be another good indicator of structural model deficits.

Reply:

We agree that, as discussed in the paper and mentioned by the Reviewer, the discharge performance criterion had a dominant influence on the multi-criteria model performance. This was especially visible when comparing different models with each other (see Figures 3 and 8 in the manuscript) and could indeed be one of the reasons for the poor simulation of the spatial pattern in the evaporation. To test this, the models were also calibrated with respect to the spatial pattern in the evaporation only. This did improve this variable, but only to a certain extent (Figure 1 below). We discussed this in Section 5 of the original manuscript, where we emphasized that it is plausible to assume that the poor simulation of the spatial pattern was more likely a result of using the same parameters within a specific HRU for all grid cells throughout the basin as also observed in previous studies (Stisen et al., 2018). We will expand the discussion of that issue in the revised manuscript and provide Figure 1 below as supplementary material to the manuscript.

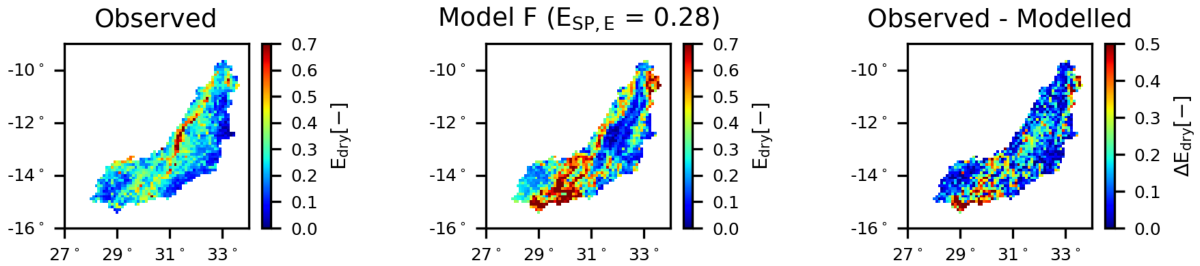

Figure 1: Spatial variability of the normalised total evaporation for Model F averaged over all days within the dry season. The left panel shows the observation according to WaPOR data, the middle panel the model result using the "optimal" parameter set with respect to the spatial pattern in the evaporation ($E_{\mathrm{SP,E}}$), and the right panel the difference between the observation and model.

Comment:

The satellite-based data product used here for calibration and validation is an actual evapotranspiration product, isn't it? I suggest to change the term evaporation to ET throughout the manuscript.

Reply:

Thank you for pointing this out. This study used the satellite product WaPOR for calibration and validation. This product describes the actual total evaporation as the sum of the individual components interception, evaporation, soil evaporation and plant transpiration (FAO, 2018). While many studies use the term "evapotranspiration" to describe the combination of different evaporation processes, other studies use the term "evaporation" as overarching term. The FAO defines "evapotranspiration" as the sum of evaporation from different surfaces and transpiration from plants. However, as argued by Savenije (2004), interception and soil evaporation, on the one hand, are functionally completely different processes than transpiration, on the other hand. While the latter is constrained by moisture in the root zone and contains a biological component of water released by stomata before the physical processes of evaporation from the surfaces of leaves occurs, interception and soil evaporation are purely physical processes. We therefore prefer to keep the term "evaporation" as an overarching term as, strictly spoken, there is no single process that can be referred to as "evapotranspiration" (Savenije, 2004; Brutsaert, 1982; 2005).

Comment:

line 114: "In this study, the long-term bias between the discharge, evaporation (WaPOR) and total water storage anomalies (GRACE) was corrected by multiplying the evaporation with a correction factor of 1.08 to close the long-term water balance." What about precipitation? Its amount is required to close the water balance.

Reply:

In general, an open long-term water balance could indeed be a result of uncertainties in precipitation, evaporation and/or discharge. As a result of limited ground observations, it was not possible to validate the satellite-based observations to correct for errors such as bias. In this study, the hydrological model and satellite-based evaporation product WaPOR used the same precipitation product CHIRPS (FAO, 2018). As a result, any bias between modelled and satellite-based evaporation cannot be a result of the precipitation (even though it could be a reason for the water balance non-closure), but can be a result of different underlying methodologies. That is why we chose to only correct the evaporation.

As simple comparison, the model was run with a random parameter set adjusting 1) the observed evaporation (factor 1.08) and 2) the precipitation only (factor 0.93). The modelled evaporation decreased slightly in Scenario 2 compared to Scenario 1 as it decreased with an average of 0.1 mm/d and a maximum of 0.5 mm/d (Figure 2 here below). The model performance with respect to the temporal variation in the evaporation was also very similar to each other with $E_{\mathrm{NS,Basin,E}}$ = 0.65 for Scenario 1 and $E_{\mathrm{NS,Basin,E}}$ = 0.66 for Scenario 2 since normalised values were used as explained in the paper.

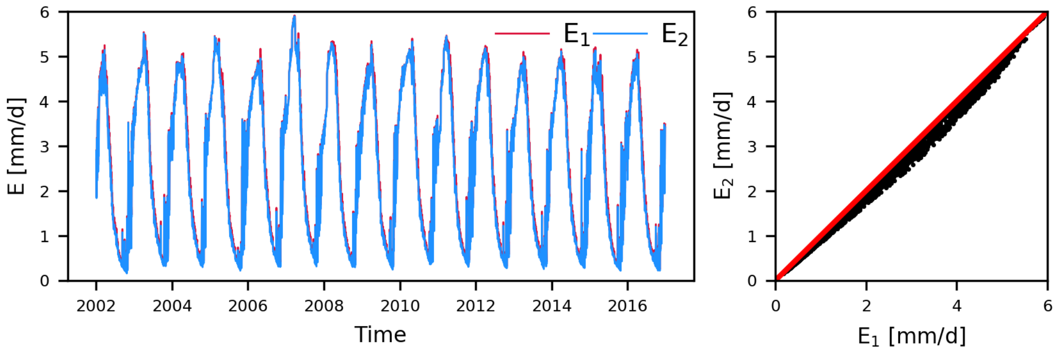

Figure 2: Modelled evaporation for one random simulation when adjusting the 1) observed evaporation (corresponding to E_1), or 2) precipitation only (corresponding to E_2) to close the long term water balance. The red line in the right panel indicates the 1:1 line.

Comment:

Figure 5, caption: "Range of model solutions for Models A to F." This should read "Model A" only.

Reply:

Thank you for pointing this out. This will be corrected in the revised version of the manuscript.

Comment:

Figure 7 and 12: "Spatial variability of the normalised total water storage anomalies for Model A averaged over all days within the dry season."

Reply:

In Figures 7 and 12, the spatial variability of the normalised total water storage anomalies was visualised for Model A. In Figure 12, also Models C and F were included. The difference between both figures is that in Figure 7, the "optimal" parameter set was selected based on discharge data only ($D_{\mathrm{E,Qcal}}$), while in Figure 12 multiple-variables were used to for this purpose ($D_{\mathrm{E,ESQcal}}$). This was indicated in the second part of the figure caption for both figures. We will further clarify this in the revised manuscript.

<u>Comment:</u>

line 403: "...since the model significantly overestimated storage anomalies in large parts of the basin." This statement can be misleading. After normalisation with Eq.33, a higher value of the model compared to GRACE indicates that the negative storage anomaly of the model is less pronounced than the one of GRACE because the averaging period considered here is the dry season?

<u>Reply:</u>

Thank you for pointing this out. This statement can indeed be confusing. With respect to the spatial pattern, spatially normalised values were compared with each other instead of absolute values. As a result, a higher normalised model value compared to the observation does not necessarily mean the actual (non-normalised) model value was also higher. However, it does mean the simulation results in this cell/region were high relative to the remaining of the basin compared to the observation.

To illustrate this, the simulated and observed dry season total water storage anomalies was visualised considering their normalised values (Figure 3 here below) and actual values (hence non-normalised, Figure 4 here below). Figure 3 shows that for Model A several cells have higher normalised values compared to the observation (e.g. the marked cell), while the actual modelled values are lower than the observation as shown in Figure 4 (please note the scale bar in Figure 4 is different for the observation and model). However, both Figures 3 and 4 show similar spatial pattern. For example in Figure 4, the marked cell in the modelled map shows a high value compared to the remaining of the basin, which was also the case in Figure 3. As a result, the spatial pattern was preserved when normalising the maps, also when calculating only with negative values as is the case when considering the total water storage map averaged over the dry season.

We will reformulate this statement to highlight this and avoid any confusion.

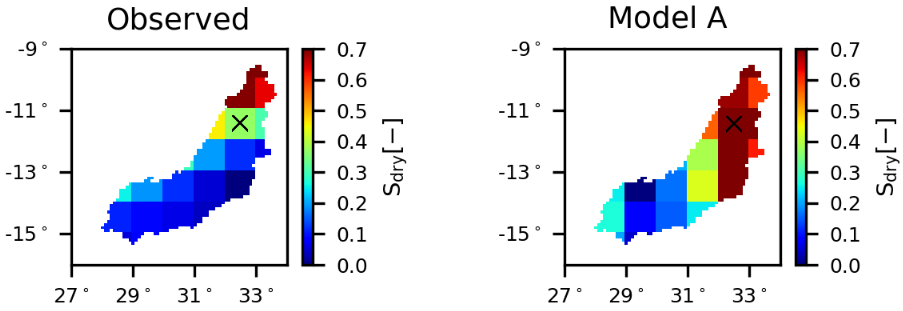

Figure 3: Spatial variability of the *normalised* total water storage anomalies for Model A averaged over all days within the dry season. The left panel shows the observation according to GRACE data, and the right panel the model result using the "optimal" parameter set with respect to discharge ($D_{E,Qcal}$).

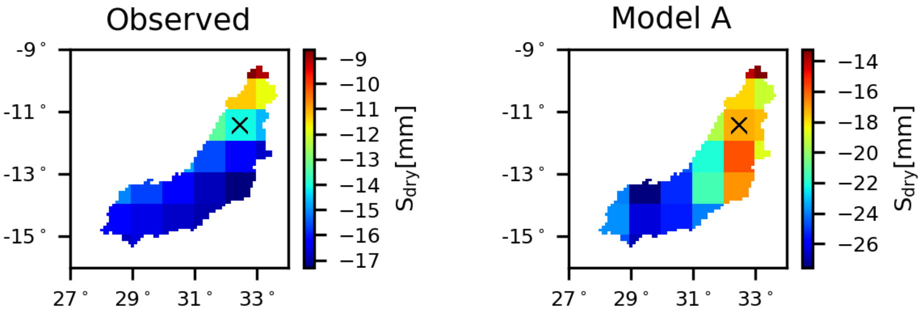

Figure 4: Spatial variability of the total water storage anomalies (*not normalised*) for Model A averaged over all days within the dry season. The left panel shows the observation according to GRACE data, and the right panel the model result using the "optimal" parameter set with respect to discharge ($D_{E,Qcal}$).

Comment:

For model calibration, a simple Monte-Carlo parameter sampling strategy is applied in spite of the fact that there are effective multi-criterial calibration methods around that can be expected to result in parameters sets with higher model performance than obtained here, such as Borg or other evolutionary algorithms. While I am not necessarily recommending to use such algorithms for the present study as its aim is rather on comparative model evaluation and development than on pure parameter optimization, the authors may explain their choice.

Reply:

Thank you for this comment. There are indeed many multi-criteria calibration methods that can be very useful to find the "optimal" parameter set and associated posterior parameter distributions. However, the goal of this study was to explore the information content of multiple variables using multiple model evaluation criteria for step-wise model structure development and calibration. For this purpose, it was important to use the same parameter sets for all models as common starting point to rule out the effect of different parameter sets. This was efficiently possible with the Monte-Carlo parameter sampling strategy, which, in addition also allowed a relatively straight-forward and intuitive interpretation and communication of the results. We will add an explanation of this choice in the revised manuscript.

References

Beven, K.: Searching for the Holy Grail of scientific hydrology: Qt=(S, R, Δt)A as closure, Hydrol. Earth Syst. Sci., 10, 609-618, 10.5194/hess-10-609-2006, 2006.

Brutsaert, W.: Evaporation into the atmosphere: Theory, history, and applications, Springer, Dordrecht, Heidelberg, London, New York, 299 pp., 1982.

Brutsaert, W.: Hydrology: An Introduction, Cambridge University Press, Cambridge, 2005.

FAO: WaPOR Database Methodology: Level 1. Remote Sensing for Water Productivity Technical Report: Methodology Series, in, FAO, Rome, 72, 2018.

Savenije, H. H. G.: The importance of interception and why we should delete the term evapotranspiration from our vocabulary, Hydrological Processes, 18, 1507-1511, 10.1002/hyp.5563, 2004.

Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., and Jensen, K. H.: Moving beyond run-off calibration—Multivariable optimization of a surface–subsurface–atmosphere model, Hydrological Processes, 32, 2654-2668, 10.1002/hyp.13177, 2018.