**Hydrology and
Earth System
Sciences**

Discussions

Interactive comment on "Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement" *by* Petra Hulsman et al.

Anonymous Referee #1

Received and published: 14 June 2020

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.

1) 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.

2) 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.

3) 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.

4) line 114: "In this study, the long-term bias between the discharge, evaporation (Wa-POR) 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.

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

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

7) 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?

8) 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.