Authors' Reply to Anonymous Referee #1

"Suitability of 17 rainfall and temperature gridded datasets for largescale hydrological modelling in West Africa" by Dembélé et al.

***Preliminary remark**: the comment numbering has been introduced by the authors for cross-referencing.*

Anonymous Referee #1

1) This is an interesting paper that is reasonably well written. Although the assessment includes a large number of datasets, the study area is relatively small, and the model is not recalibrated for each variable, which has led to some questionable conclusions.

***Response:** We thank the referee#1 for the positive overall appreciation of our work. As it can be read from the title, our study focuses on the poorly studied region of West Africa, when it comes to hydrological modelling in general and hydrological evaluation of meteorological datasets in particular. Accordingly, our contribution clearly represents added-value, in terms of regional hydrology as well as in terms of hydrological modelling of semi-arid areas. In fact, independent regional evaluation of globally and regionally available datasets are of key importance for hydrology as they can provide new insights that might not be fully highlighted in global studies.*

Regarding the general critics on the model calibration, we would like to refer the reader to our detailed answer to comment 4. Below we provide more details and answers to the detailed comments of this reviewer.

2) You state that "rainfall datasets have contrasting performances across the four climatic zones present in the VRB, suggesting that, in general, basin-wide hydrological model performance might be misleading and invalid for a smaller spatial domain." What makes you think that your results, which also represent a relatively small spatial domain, are not "misleading and invalid" as well?

***Response:** We thank the referee#1 for pointing out this potentially misleading statement. We agree with this referee that our statement has some strong wordings, which have led to a different interpretation by the referee#1 than what we intended to say. Our original idea is that the overall model performance in the entire, relatively large modelling domain (from a catchment hydrology perspective, the Volta River Basin, VRB, is indeed a large domain) might not be representative for all subdomains. This is especially the case if the modelling domain extends over multiple climatic zones as in the VRB case. For instance, in global studies, the overall global performance of a rainfall dataset is likely different from its performance in sub-regions such as West Africa (e.g. see Figure 3 in Beck et al. 2017b). Therefore, by "smaller spatial domain" we meant a portion of a large domain under evaluation. We will correct this statement to avoid any misunderstanding in the revised manuscript.*

3) It is stated that "the results can be considered valid for West Africa and regions with similar hydroclimatic and physical features" which is highly speculative and likely not true given the variation in precipitation dataset performance and gauge network density. To improve the generalizability of the results, the assessment should be expanded to other regions across Africa or the globe.

Alternatively, the abstract and discussion should clearly state that the conclusions and the performance ranking of the datasets are not representative of other regions.

Response: *We agree with the referee#1 that this isolated statement can be interpreted as speculative. However, we did not want to imply any certainty but the possibility that the results might be transferable to other places. This is expressed in the sentence following the aforementioned sentence: "A wider generalization of the findings should be done with caution and after repeating similar evaluation studies in other places". Also, we did not intend to generalize our findings to other regions, which is very clear from the mentions "West Africa" in the title, and "Volta River Basin (VRB) in West Africa" in the very first sentence of our manuscript.*

We agree that we should not have mentioned transferability to other similar climates outside Africa since the performance of any remote sensing-based meteorological data set for hydrological modelling varies across the globe due to many other factors not only related to regional aridity.

To avoid ambiguities, the statements will be reformulated as follows in the discussion: "The results are primarily valid for the study region in West Africa, while a wider generalization of the findings should be done with caution and after repeating similar evaluation studies in other places".

4) The soil moisture, terrestrial water storage, and actual evaporation assessments were carried out without recalibrating the model and therefore the results for these variables are subject to substantial uncertainty. This is supported by the fact that MSWEP, which was used to force GLEAM, does not exhibit good actual evaporation scores. The model should be recalibrated for each variable.

Response:

Performance of MSWEP: In our opinion, MSWEP has very good scores for modelled actual evaporation compared to GLEAM as it always exceeds a Pearson correlation coefficient (r) of 0.9 (Figure 8), with an average r=0.94 for the entire VRB (Appendix A3), and it has the highest spatial pattern score (Esp=0.26) among all the rainfall datasets (Appendix A3).

Model recalibration: We would like to emphasize here that the model is indeed recalibrated for each meteorological input product combination (i.e. rainfall and temperature), but it is recalibrated with streamflow (Q) only and not with soil moisture (Su), terrestrial water storage (St), and actual evaporation (Ea). The logic behind this approach is twofold:
i) We would like to know how well the model performs in combination with the different input variables; we therefore use Su, St and Ea as evaluation variables.
ii) Further calibrating our model with Su, St, and Ea would lead to additional model improvement due to the information content of these variables as demonstrated by Dembélé et al. (2020). In this case, it becomes difficult to disentangle the contribution of the rainfall datasets and the contribution of the calibration variables (Su, St, and Ea) to the overall model performance. Calibrating the model on one reference output variable (in-situ streamflow) and evaluating it against other output variables remains in our view a powerful method to assess the usefulness of a meteorological input dataset for hydrological modelling. By calibrating on streamflow, we give each meteorological data set "a chance" to perform as well as possible for streamflow; we then further discriminate between the usefulness of the input variables for hydrological modelling by assessing whether they can do a good job for streamflow and Su, St and Ea simultaneously. The "dream" input variable should indeed perform well for all variables if only calibrated on one.

We would like to emphasize here, that the evaluation of Su, St, and Ea is not done with the absolute values (i.e. raw data) of the satellite products, but rather we evaluate their temporal dynamics and spatial patterns using bias-insensitive metrics. Therefore, we substantially mitigate uncertainties that might arise from the assessment of these variables when using their absolute values (Dembélé et al. 2020; Nijzink et al., 2018; Mendiguren et al., 2017; Wambura et al., 2018). We had already discussed our choice for the Q-only calibration and its limitations at lines 456-458 and the potential uncertainties related to the satellite datasets used for evaluation at lines 441-443. However, we will now make the choice of the Q-only calibration clearer by adding the following in the discussion: "The model is calibrated only on Q data despite the known limitations of the Q-only calibration (Demirel et al., 2018). However, calibrating the model on additional variables would result in additional improvement in model performance that would not be separable from the contribution of the input datasets to the model performance. Therefore, regarding the goal of this study, the Q-only calibration was the best option to obtain the impact of various meteorological forcing datasets on the plausibility of hydrological processes."

5) The word "gauge" is not used in the abstract and the datasets are only classified as either satellite or reanalysis. However, the amount of gauge data incorporated in the datasets may well be the overriding factor in determining the performance, given the good performance of TAMSAT and CHIRPS in terms of streamflow.

Response: *In the abstract, we will modify the statement "Seventeen precipitation products based on satellite data (…)" into "Seventeen precipitation products based essentially on gauge-corrected satellite data (…)". Moreover, Table 1 provide information on rainfall datasets developed with gauge data.*

6) Figures 7 and 10 are impossible to interpret, way too much information. Should be condensed.

Response: We agree with referee#1 that Figure 7 and 10 contain a lot of information. We will replace Figure 7 and 10 by Figure S30 and S48, respectively. Thereby, only showing the model performance for the entire VRB, while the performance for the four climatic zones will be moved to the supplementary materials.

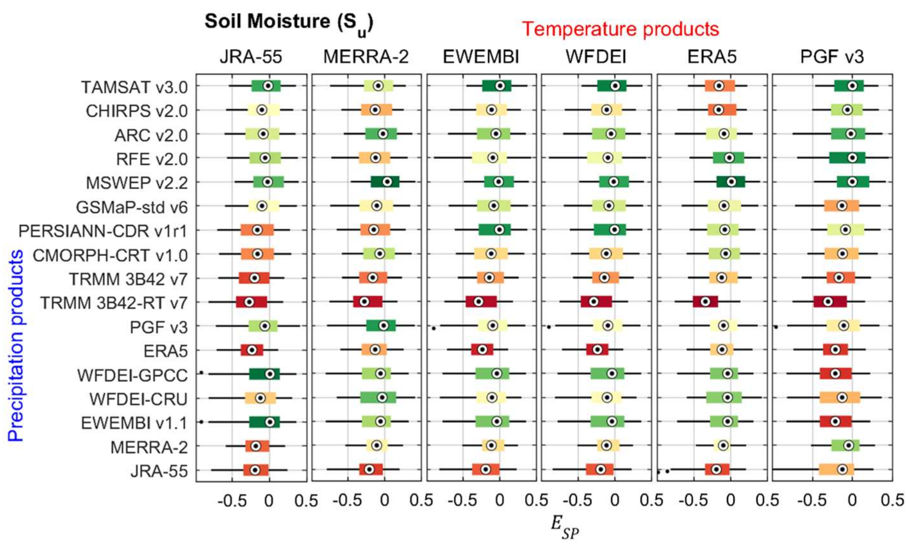

Figure S30. Spatial pattern efficiency (E_{SP}) of soil moisture (S_u) over the entire simulation period (2003-2012) for the Volta River basin (VRB), using different combinations of precipitation and temperature products for hydrological modelling. Each boxplot has 120 values corresponding to the number of months. The boxplots are colored from the best (green) to the worst performance (red) based on the median value.

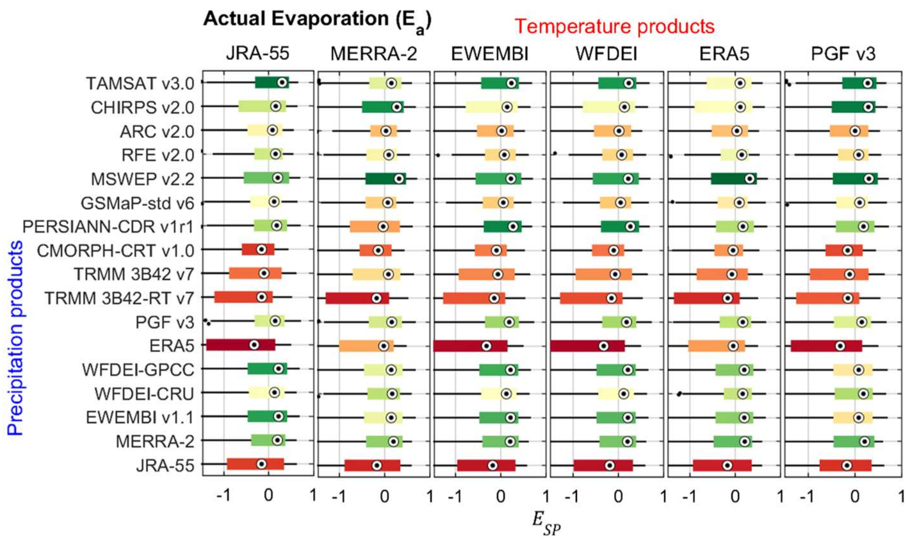

Figure S48. Spatial pattern efficiency (E_{SP}) of actual evaporation (E_a) over the entire simulation period (2003-2012) for the Volta River basin (VRB), using different combinations of precipitation and temperature products for hydrological modelling. Each boxplot has 120 values corresponding to the number of months. The boxplots are colored from the best (green) to the worst performance (red) based on the median value.

References

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrol Earth Syst Sc, 21, 6201-6217, https://doi.org/10.5194/hess-21-6201-2017, 2017b.

Dembélé, M., Hrachowitz, M., Savenije, H. H., Mariéthoz, G., and Schaefli, B.: Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite datasets, Water Resources Research, https://doi.org/10.1029/2019WR026085, 2020.

Demirel, M. C., J. Mai, G. Mendiguren, J. Koch, L. Samaniego, and S. Stisen (2018), Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, Hydrology and Earth System Sciences, 22(2), 1299-1315, https://doi.org/10.5194/hess-22-1299-2018.

Mendiguren, G., J. Koch, and S. Stisen (2017), Spatial pattern evaluation of a calibrated national hydrological model - a remote-sensing-based diagnostic approach, Hydrology and Earth System Sciences, 21(12), 5987-6005, https://doi.org/10.5194/hess-21-5987-2017.

Nijzink, R. C., et al. (2018), Constraining Conceptual Hydrological Models With Multiple Information Sources, Water Resources Research, 54(10), 8332-8362, https://doi.org/10.1029/2017wr021895.

Wambura, F. J., O. Dietrich, and G. Lischeid (2018), Improving a distributed hydrological model using evapotranspiration-related boundary conditions as additional constraints in a data-scarce river basin, Hydrological processes, 32(6), 759-775, https://doi.org/10.1002/hyp.11453.