Authors' Reply to Referee #2 (Dr. Nadav Peleg)

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

Review (hess-2020-68)
In their paper, Dembélé et al. explore the suitability of combining time series of rainfall and temperature from different climate products as inputs into a hydrological model. The manuscript is well structured and written, methods are robust and results are presented adequately. The research question of the possibility of combining gridded climate variables from different sources to simulate various hydrological components is relevant and timely, and I believe will be of interest for the readers of HESS. Nevertheless, I have a few comments and suggestions for the authors to consider before I can recommend the paper for publication.

Sincerely,

Nadav Peleg

Response: *We thank the referee#2 for the positive overall appreciation of our work. Below we provide answers to the referee's comments.*

Major comments
1. I found one-step in the methodology (i.e. as presented in Figure 1) to be missing. I think it will be meaningful to know how the climate variables (rainfall, temperature) from each climate products are ranked in comparison to observed data (i.e. from ground stations) before ranking the 102 input combinations based on various hydrological components. I think this step is critical to understand the presented results. For example, JRA-55 and ERA5 yield poor correlation with Ea (Figure 8), but isn't this because they are poorly reproducing the rainfall statistics over the VRB? GSMaP-std V6 reproduces well the streamflow (Figure 3), St (Figure 4), Su (Figure 5) and Ea (Figure 8) – will this product be ranked #1 when compared to ground stations? I assume there will be a high correlation between the ranks emerging from the comparison to ground stations and hydrological outputs from the model. If this case, wouldn't it be sufficient to evaluate the best products to use in hydrological simulations simply by comparing them to the few climate stations that are available in the catchment of interest or a nearby area? This is a point for discussion.

Response:

Comparison with ground observations

We agree with the referee that knowing the performance of the meteorological datasets in comparison with ground measurement could be an interesting starting point. However, it is noteworthy that the Volta River basin (VRB) in West Africa is a data scarce region, not like other places in Europe and USA (e.g. Beck et al., 2019a) where a large amount of ground measurements is widely and freely accessible. The few datasets collected by local organizations in the VRB are not easily accessible due to the transboundary nature of the basin that is shared among six countries. It took us one year to obtain streamflow data, which was further subject to a thorough gap-filling and quality control of time series (Dembélé et al., 2019).

The VRB region has a low density of meteorological stations (see Figure 1 of Dembélé and Zwart 2016; and Figure 1 of Satgé et al., 2020). A thorough evaluation of satellite/reanalysis datasets with ground measurements in the VRB cannot be limited to a few stations because the basin is about 415,600 km^2 (ten times the size of Switzerland), with a unique and complex climate (see Section 2.4 Study Area), and a strong spatial variability of rainfall.

Even in case of ground measurement availability, the validity of point-to-pixel comparison is questionable (e.g. JRA-55 is 1.25°, and ERA5 is 0.25°) because the gauge measurement will hardly represent the spatial variability of rainfall in a pixel. Moreover, the rainfall datasets used in our study are essentially gauge-corrected data. Therefore, a robust ground evaluation would require independent ground measurements that are not used in the development of the rainfall datasets (Beck et al., 2019a), which is a luxury in West Africa.

<u>Validity of ground evaluation for hydrological modelling</u>

The skill of a product in reproducing well ground measurement under a point-to-pixel evaluation does not necessarily guarantee its high performance for hydrological modelling, mainly in complex hydroclimatic environments such as the VRB. The performance of isolated pixels might not be representative of all pixels. Usually, hydrological modelling is undertaken at daily or higher temporal resolution. However, mismatches between gauge and satellite reporting times are a major issue in ground evaluation (Beck et al., 2019a). This is confirmed by the substantial increase in the evaluation performance of rainfall datasets from daily to monthly time scale (Dembélé and Zwart (2016); see Figure 3 vs. Figure 8 of Satgé et al. (2020)).

We will add the following to our discussions: "When comparing the results of this study to the findings of Satgé et al. (2020) based on a point-to-pixel evaluation of gridded rainfall datasets in West Africa, it is noticeable that the ground evaluation might lead to different results as compared to the hydrological evaluation as adopted in the current study. The skill of a rainfall product in reproducing well ground measurement under a point-to-pixel evaluation does not necessarily correlate with its performance for hydrological modelling, mainly in large and complex hydroclimatic environments such as the VRB. Therefore, ground evaluation it is not always a needful step before hydrological evaluation of gridded rainfall datasets."

2. The modelling experiment includes 6 years for model calibration and 4 years for model evaluation. These are very short periods, not necessarily representing well the natural climatic and hydrological variability and not necessarily guarantying a successful calibration of the hydrological model parameters. First, I suggest demonstrating with a simple graph (can be presented as SI) that the natural variability is somehow represented in your 10-year data. Second, consider adding a short discussion regarding the sensitivity (quantified) of the hydrological model parameters to the short period that is used for the model training.

<u>Response</u>:

<u>Length of the calibration and simulation period</u>

We agree that the modelling period of 10 years, which includes 6 years for calibration and 4 years for evaluation, might not seem very long, but is long enough to obtain a well calibrated model in our case, as previously demonstrated by Dembélé et al., (2020). Moreover, a 3-year model warm up period (2000-

2002) precedes the calibration period. The choice of the modelling period is constrained by the availability and the quality of the in-situ streamflow measurement in the data-scarce VRB (Dembélé et al., 2019).

Moreover, it is important to stress that we adopt a daily streamflow calibration, which means a time series of 2192 time steps to simulate and match for each of the 11 gauging stations during the 6-year model calibration period (2003-2008), or 3653 time steps for the 10-year simulation period (2003-2012). In our opinion, this is a robust model calibration approach, additionally supported by the fact that we adopt a multi-site calibration simultaneously at 11 streamflow gauging points located in very distinct hydroclimatic zones within the basin (see Figure 2). It is worth mentioning that the computational cost for each of the 102 input data combinations is about 6 days for 4000 parameter iterations during the model calibration on a computer Intel Xeon Processor E5-2697 v3 with 64 GB of RAM.

Natural variability of streamflow

We thank the referee#2 for this important comment on natural variability that was not appropriately discussed. Natural variability of daily streamflow can be observed at each of the 11 streamflow gauging sites used in this study, and inter-site variability of streamflow can be observed as well for the 10-year period (2003-2012). As it can be seen in Figure R1 below, the modelling period covers years with considerably different streamflow volumes during the wet season and with considerably different peak discharges, ranging e.g. for station 4 from 250 m3/s (year 2011) to 900 m3/s (year 2003). In general, years 2004, 2005 and 2009 can be considered as dry while 2003, 2006, and 2010 are wet for station 2 and 4, which have low flows as compared to the station 11.

Similar figures showing the hydrographs of all the eleven stations will be added to the supplementary materials and indexed in the text of the revised manuscript.

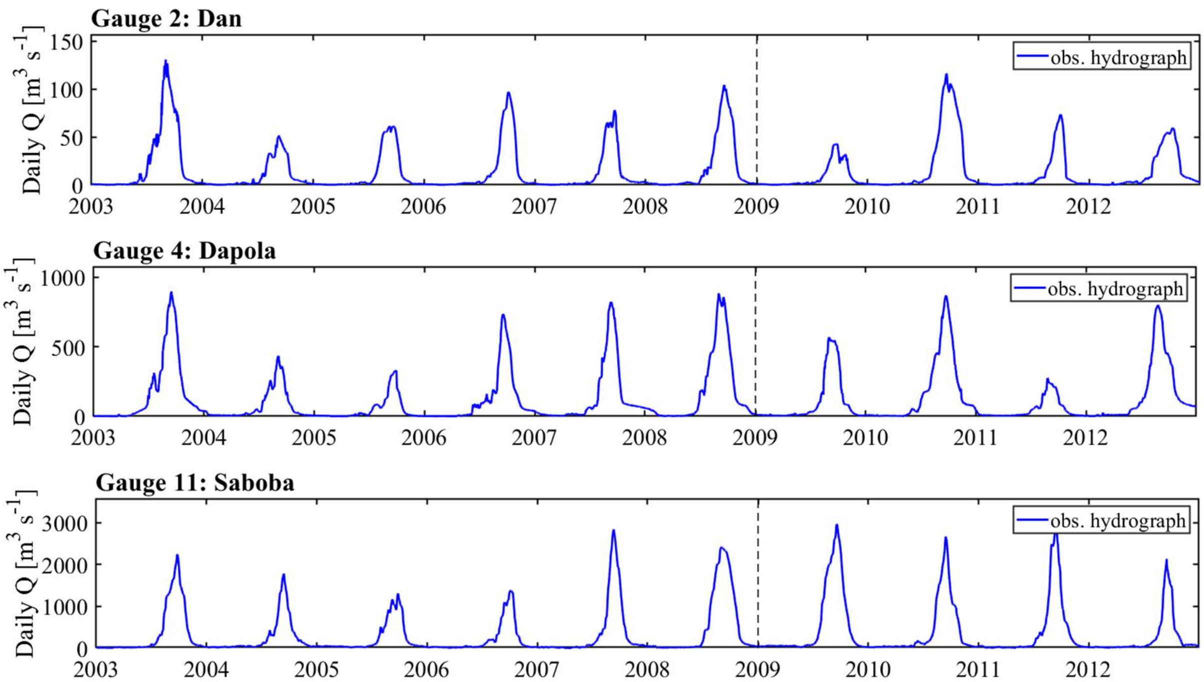

Figure R1: Hydrographs at three different gauging stations in the VRB during the modelling period comprised of the calibration period (2003-2008) and the evaluation period (2009-2012)

<u>*Natural variability of meteorological datasets*</u>

Natural variability can also be observed in the rainfall datasets as shown below in Figures R2-R3. It can be seen that rainfall varies both in time and space across different climatic zones in the VRB, which makes it an interesting case study for rainfall evaluation.

These figures will be added to the supplementary materials and indexed in the text of the revised manuscript.

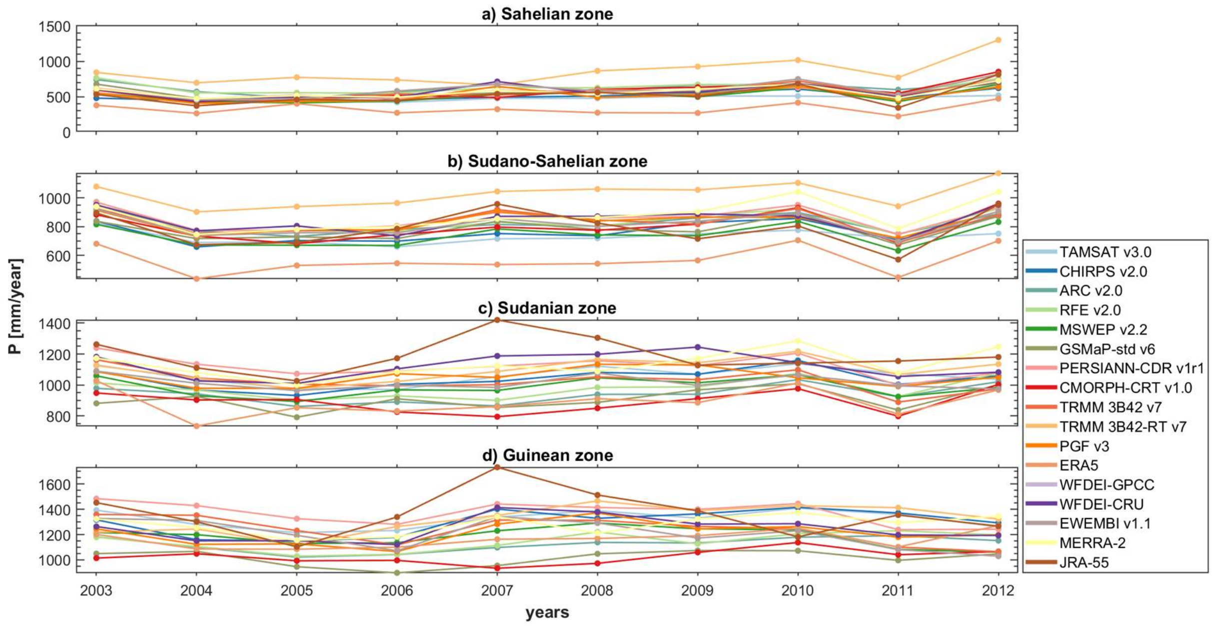

Figure R2: Annual total rainfall for 17 datasets for different climatic zones in the VRB

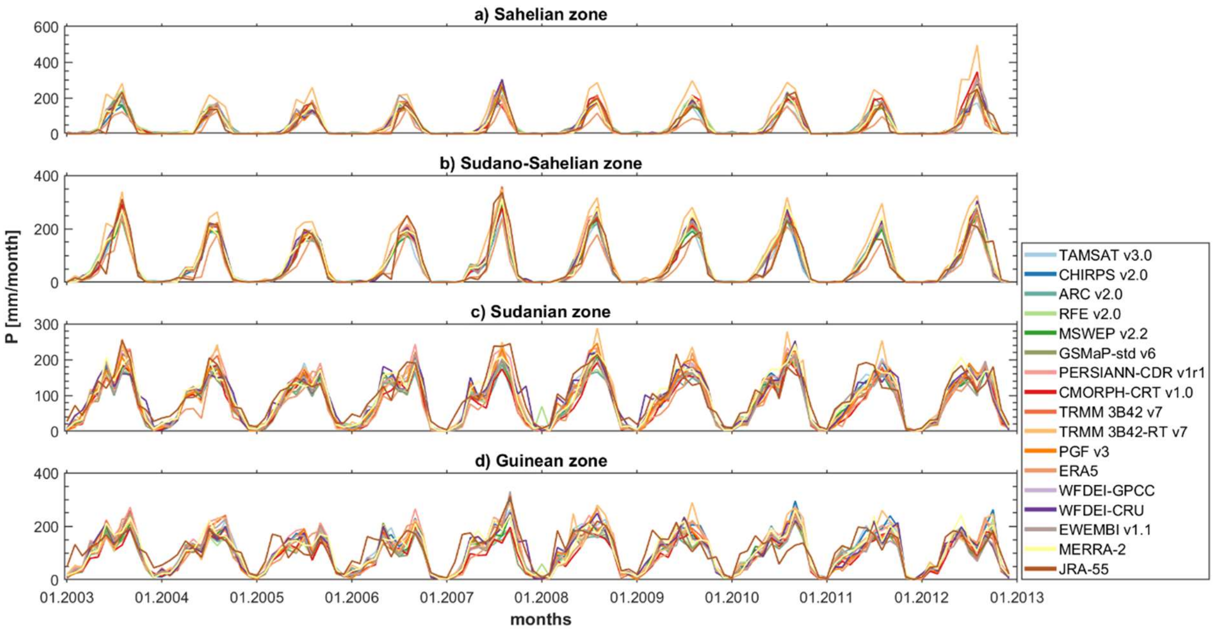

Figure R3: Monthly total rainfall for 17 datasets for different climatic zones in the VRB

Sensitivity of model parameters

In the supplementary materials (Figure S65 at Section 9.2 on P61), we provide a figure that shows the distribution of each of the 36 global model parameters and their sensitivity (i.e. second-order coefficient of variation) to different input meteorological data. It can be seen that most of the model parameters vary considerably as a response to the change of rainfall and temperature data.

The following will be added to the discussion: "(…). Moreover, it can be noticed that most of the model parameters are sensitive to the change in meteorological input datasets (Figure S65)."

Minor comments

1. Usually, when considering using gridded climate variables from climate re-analysis/other products as inputs into hydrological models the following steps are taken: (i) computing the skills (i.e. temporal dynamics, magnitude, and occurrence) of the climate variables in comparison to observed data; (ii) choosing the (individual) climate product with the best skill to use; and (iii) performing a bias correction to the climate variables to improve the fit to the observed data. I am missing a paragraph in the introduction/discussion explaining why not simply following this practice which should improve the hydrological outputs from the model.

Response: *The approach proposed by the referee#2 is usually applied for hydrological climate change impact studies, where climate projection data known to be biased are first evaluated and corrected with observations. In our manuscript, we have described at lines 78-85 the usually adopted approaches for the evaluation of gridded (satellite and reanalysis) datasets as follows:*
"The errors quantification of SRPs and reanalysis products is usually done by comparing them with in-situ measurements (e.g. Dembélé and Zwart, 2016;Thiemig et al., 2012;Beck et al., 2019a;Caroletti et al., 2019;Satgé et al., 2020), or by assessing their reliability as forcing for hydrological models (e.g.Duethmann et al., 2013;Pan et al., 2010;Nkiaka et al., 2017). Other evaluation approaches include triple collocation, which is a technique that estimates the variance of unknown errors of three independent variables without a reference or observed variable (e.g. Massari et al., 2017;Alemohammad et al., 2015;McColl et al., 2014;Roebeling et al., 2012). Compared to the ground-truthing approach, the hydrological evaluation approach has received limited attention (Camici et al., 2018;Poméon et al., 2017)."

Among those approaches, we adopted the hydrological evaluation, which consists in assessing the reliability of the gridded datasets in reproducing plausible spatiotemporal patterns of hydrological processes when used as input to a model, knowing that they might still present some discrepancies with ground measurements. This approach is particularly interesting in data scarce regions where ground evaluation is challenging or impossible. It is important to mention here that the gridded datasets that we are evaluating in our study are essentially gauge-corrected datasets as mentioned at lines 160-161, also see Table 1. In this case, the datasets are already bias-corrected.

We will make this clearer in the abstract by mentioning the use of gauge-corrected datasets in our study. Therefore, as also requested by the referee#1, the statement "Seventeen precipitation products based on satellite data (…)" will be replaced by "Seventeen precipitation products based essentially on gauge-corrected satellite data (…)". Moreover, Table 1 provides information on rainfall datasets developed with gauge data.

2. Results (Figure 3, for example). 22 values are used to represent the combined performance for the calibration and evaluation periods. This is not clear to me. Why not using a single Ekg value for the entire simulation period (merging the calibration and validation periods to a single period) for each gauge, i.e. 11 values in total per combination of temperature and precipitation product? What is the logic in separating the Ekg values to calibration and validation periods?

Response: *The decision for using 22 values of* E_{KG} *(11 for calibration + 11 for evaluation) was based on the necessity to have enough elements for plotting the boxplots. For simplicity in reporting, a new Figure 3 will be provided only showing the median* E_{KG} *of the entire simulation period, similarly to Figure 4, 5 and 8.*

3. Table 1. I suggest adding in the table additional column indicating if the product refers to rainfall, temperature or both. Also, please double-check the space-time resolutions reporter in the table. I think that the CMORPH-CRT product, for example, has a resolution of 8-km and 30-min.

Response: *We thank the referee for the suggestion, which will be considered in the revised manuscript. We are aware that different versions of the datasets exist, so that we have carefully mentioned in the caption of Table 1 that the information provided refer to the version of the datasets we have used. The provided information for CMORPH-CRT is correct, and the data was accessed from this web link:* ftp://ftp.cpc.ncep.noaa.gov/precip/CMORPH_V1.0/CRT/0.25deg-DLY_00Z/

4. The use of second-order CV is interesting, I do not recall seeing it in the context of hydrological statistics. Why use it and not simply using Pearson's CV skill? A sentence explaining the motivation is needed.

Response: *We agree with the referee that the use of the second-order CV is uncommon. In the revised manuscript, we will add the reasons of it use instead of the classical Pearson's CV, which has major limitations that are comprehensively described by Kvålseth (2017).*

5. Figures 7 and 10. Too many box-plots are presented. Perhaps present only the median (avoid using box-plots) to compare between products and climatic zones. This will considerably reduce the size and information plotted.

Response: *We agree with the referee#2 that Figure 7 and 10 contain a lot of information. As also responded to the referee#1 (comment #6), 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.*

6. Generalization of the results. In lines 437-438 you mentioned that: "The results can be considered valid for West Africa and regions with similar hydroclimatic and physical features. A wider generalization of the findings should be done with caution and after repeating similar evaluation studies in other places". I do not think that you can generalize the results - they are likely to differ between locations as the quality of climate variables from different climate products differ between locations. In my view, the key message of your paper is that for each large catchment you should consider multiple sources of climate data to find the climate variables combination that is suitable for your region. The VRB is simply a case study used to demonstrate this point.

***<u>Response</u>:** We agree with the referee#2 and we would like to stress that we did not intend to generalize our results as we carefully draw the reader's attention on the necessity to repeat the same experiment in other regions. As also responded to the referee#1 (comment #3), and to avoid ambiguities, the statement will be modified as follows: "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".*

References

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I., Huffman, G. J., Adler, R. F., and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrol Earth Syst Sc, 23, 207-224, https://doi.org/10.5194/hess-23-207-2019, 2019a.

Dembélé, M., and Zwart, S. J.: Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa, Int J Remote Sens, 37, 3995-4014, https://doi.org/10.1080/01431161.2016.1207258, 2016.

Dembélé, M., Oriani, F., Tumbulto, J., Mariéthoz, G., and Schaefli, B.: Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings, Journal of Hydrology, 569, 573-586, https://doi.org/10.1016/j.jhydrol.2018.11.076, 2019.

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, e2019WR026085, https://doi.org/10.1029/2019WR026085, 2020a.

Kvålseth, T. O.: Coefficient of variation: the second-order alternative, Journal of Applied Statistics, 44, 402-415, https://doi.org/10.1080/02664763.2016.1174195, 2017.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10(3), 282-290, https://doi.org/10.1016/0022-1694(70)90255-6.

Satgé, F., Defrance, D., Sultan, B., Bonnet, M.-P., Seyler, F., Rouché, N., Pierron, F., and Paturel, J.-E.: Evaluation of 23 gridded precipitation datasets across West Africa, Journal of Hydrology, 581, 124412, https://doi.org/10.1016/j.jhydrol.2019.124412, 2020.