Discussion comment 1

In the manuscript titled “MSWEP: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data” authors have merged several satellite and reanalysis only precipitation products. This merged product is later validated using precipitation data sets that are not used in the merging process and using HBV hydrological model outputs. Results of both validation efforts show the merged product is on average superior to input products. The idea of merging different products to obtain a better one sounds trivial, yet in this case it results in a product that may have large application areas. The topic is relevant to HESS Journal, and both the methodology and the validation efforts sound good. I recommend the study to be published after correction of some points.

We would like to sincerely thank the reviewer for their positive remarks and thorough review.

1.a) I found the methodology section related with HBV model rather short, it would be idea if it is expanded. I guess NSE is calculated between observed and simulated Q values, but I couldn’t find this info written explicitly. I am not very familiar with the HBV model, so the parameter calibration part seems not clear to me (e.g., how did authors implemented the calibration, using a particular software? Running the model with different combinations of parameters sampled randomly from their defined range in Table 3?)

We appreciate the suggestion. We now explicitly mention in the methodology and in the captions of Figure 8–10 that the NSE is calculated using simulated and observed *Q* time series. The discussion m/s already explicitly mentions the software with which we implemented the calibration including a citation (the Distributed Evolutionary Algorithms in Python–DEAP–toolkit; Fortin et al., 2012; see page 11 lines 32–33). The discussion m/s also already explicitly mentions the parameter optimization scheme which we employed on page 11 line 32 (the (μ+λ) evolutionary algorithm). We added an additional reference for readers interested in more details (Ashlock, 2010). Note that the software and optimization scheme can be used with any hydrological model—it is not part of the HBV model or provided with it.

Ashlock, D. (2010): Evolutionary computation for modeling and optimization, Springer Publishing Company, pp. 572.

1.b) Did HBV calibration and validation efforts use the same runoff (Q) data? If they are the same, then it is very likely that the calibration might fit Q observations too closely (which is a particular advantage on MSWEP compared to other products).

HBV was re-calibrated for each precipitation product independently, thus MSWEP was not given any unfair advantage. There was in fact no validation exercise—our objective was to quantify the information content of each product and we are therefore only interested in the calibration NSE scores, which we report in the m/s.

2) Do Reanalysis data have 5m wind dataset instead of converting 80m height to 5m using some wind-profile relation?

Unfortunately, the Vaisala data are only available at 80-m height since they are originally intended for wind-power related applications. Although other reanalysis datasets exist that do provide wind speed estimates at more appropriate heights, they tend to have a much lower spatial resolution and consequently fail to provide realistic wind speed estimates in many mountainous regions.

3) E = P – Q. If E is only evaporation (line 6, page 7), then what happens to transpiration component?

Our definition of evaporation includes transpiration (please see Savenije, 2004, for a discussion on “evaporation” versus “evapotranspiration”).

Savenije, H. H. G. (2004), The importance of interception and why we should delete the term evapotranspiration from our vocabulary. *Hydrol. Process.*, 18: 1507–1511. doi: 10.1002/hyp.5563.

4) Why normalize absolute bias? The unit of the bias is very important as well.. It would be complementary with RMSE (i.e., the random components can be calculated if non-normalized bias and RMSE are known).

We did not normalize the bias (*B*). Perhaps the reviewer is asking why we took the absolute value of *B*? This is because, in this study, we are not interested in whether the products generally overestimate or underestimate. Rather, we are interested in how far off the values are on average, or in other words, how well the product performs on average. RMSE computes the difference for each time step and is thus completely unrelated to *B*, which averages each time series prior to calculating the difference.

5) Figure 6a, Long-term average weights would have been more meaningful rather than arbitrarily chosen single day.

We appreciate the suggestion, and have given this some thought, but in the end we decided to only show an example for a single day. This is because, if we would show the long-term average, we would have to do this separately for the pre- and the post-TRMM eras, and for the monthly, daily, and 3-hourly time scales. This would result in 18 figures spanning two pages, which would detract from the main message of the m/s. However, based on this comment we have decided to release the weight maps as part of the NetCDF files so people can use them in their research and derive any spatial-temporal average they like .

6) It is not really clear to me why reanalysis HBV performance is much worse than MSWEP given there is only minor difference between them in terms of accuracy of P (Figure 7)?

The distribution of the FLUXNET stations is completely different from that of the streamflow gauges, leading to very different performance scores. Notably, there is a lack of FLUXNET stations in the tropics, where reanalysis-based products tend to perform poorly.

7) NSE increases with increasing distance for Reanalysis? How come farther away gauges give more reliable precipitation information compared to closer gauges? I might be missing something simple.

Good question. This is due to the uneven distribution of the streamflow gauges around the globe, as explained in the discussion m/s (page 17 lines 20–22): "Contrary to expectations, the median NSE scores increase with increasing distance to the closest P gauge for several of the P datasets, which is primarily because the more sparsely gauged groups contain less (semi-)arid catchments, which tend to exhibit lower NSE scores."

Minor
- Page 3, line 8, "These datasets have . . .".

Thanks for the suggestion. However, we are referring to precipitation products in general, not specifically to the aforementioned products. We therefore prefer to keep "The datasets have …".

- Table 1, CMORPH does not use gauge data, why it is included in "Gauge, satellite" row? There is another row specifically dedicated to "satellite" (products 19 and 20).

As stated in the caption of Table 1, we list for each dataset only the "best" variant. So for CMORPH we refer to the variant that incorporates gauge data (which does in fact exist).

- Consider using the word "using" instead of "in turn"/"in turn with". It is very confusing.

Changed. Thank you for the suggestion.

- Figs. 8-11 captions should include very brief info about the parameter used in NSE calculation (i.e., Q).

We have added to each figure's caption that "The NSE scores have been computed between simulated and observed *Q* time series."