Short comment

L. Brocca

We would like to thank Dr. Brocca for the valuable comment and in the interest of transparency we highlight that we have interacted with Dr. Brocca beforehand, which led to the removal of the SM2RAIN-SMAP data set from the evaluation.

I was pleased to read the paper by Beck et al. who performed a comprehensive assessment of multiple precipitation datasets over Contiguous United States (CONUS). I believe the paper is a valuable contribution. However, by reading the paper two questions raised to my mind. I believe the authors might want to address these two questions in their paper:

1) What is the value of using the Kling-Gupta Efficiency (KGE) for assessing the performance of precipitation datasets? Is it suitable for determining the products performance for applications (e.g., flood prediction)?

As mentioned in the paper, the KGE is an objective and widely used performance metric that combines three fully independent performance metrics: correlation, bias, and variability ratio. The metric thus evaluates the most important aspects of the precipitation time series and is therefore *"suitable for determining the products performance for applications (e.g., flood prediction)".* However, it is important to present the results for all three components in addition to the KGE scores, as we have done in our paper.

2) Are the results obtained over CONUS representative of other regions? Can we generalize the obtained results?

The comprehensive global-scale evaluation by Beck et al. (2017) showed similar results between Europe and the CONUS, suggesting that the results of our new study are generalizable to Europe. However, some of the datasets used in the comparison (i.e., ERA5) use additional data in their assimilation over the CONUS and hence transferability of the results may be challenging. Therefore, we emphasize the need for more research to verify and supplement the present findings in the last paragraph of the Conclusions.

To answer these questions, and following the final suggestions of the authors "Similar evaluations should be carried out in other regions with ground radar networks (e.g., Europe) to verify and supplement the present findings.", we tested three different satellite-based products in Europe:

a) SM2RAIN-ASCAT dataset, i.e., a recent version of an SM2RAIN-based dataset based on the application of SM2RAIN to ASCAT soil moisture product (Brocca et al.,2017) (apologize for self-citations). This dataset is similar to SM2RAIN-CCI V2 dataset used in Beck et al.

b) TMPA, the real time version of 3B42RT, i.e., TMPA-3B42RT V7 in Beck et al.

c) CMORPH, the real time version of CMORPH, i.e., CMORPH V1.0 in Beck et al.

We have considered 646 basins in Europe, and by following the same approach proposed in Camici et al., 2018, we tested the three satellite-based products (uncorrected) against ground-based precipitation (E-OBS dataset as reference, Haylock et al., 2008) for rainfall dataset assessment, and against observed discharge observations through the application of rainfall-runoff modelling.

The figure at the end of the document shows the results, in the top, for rainfall assessment by using different performance scores (KGE, R: correlation, BIAS, REL.VAR.: relative variability, RMSE: root mean square error, ubRMSE: unbiased RMSE), and in the bottom for discharge assessment by using the KGE as target score. Each dot represents a basin in which the comparison between satellite products and E-OBS is performed for rainfall assessment (basin average rainfall), and the comparison between simulated (through rainfall-runoff modelling and the three satellite rainfall datasets as input) and observed discharge is carried out for discharge assessment. The title of each plot shows the median value of the score.

The results shown in the figure are quite interesting and illustrative of the problem in selecting a score for rainfall datasets assessment. We suppose as target the results in terms of KGE for discharge assessment shown in the bottom row.

Firstly, we underline that the results in terms of KGE for rainfall assessment (first row in the top) are not representative of the results in terms of KGE for discharge assessment. Also the use of other rainfall scores might be not suitable, with the better performance in terms of relative rankings between the products obtained by using BIAS, RMSE and ubRMSE. However, in terms of spatial assessment, each score applied to rainfall assessment seems to be not representative of discharge performances.

First, in his discharge assessment, Dr. Brocca relies on a single performance metric (KGE). We want to caution against relying on a single performance metric as this leaves room for speculation when interpreting the results.

Second, the transformation from rainfall to runoff is a complex and highly non-linear process. All rainfall-runoff models (both physical and conceptual) are simplified representations of reality and therefore require calibration. Evaporation can easily compensate for many kinds of errors in precipitation, particularly since it is such a poorly observed variable. Channel routing may also mask certain errors in precipitation, due to its smoothing effect. It is possible that the employed rainfall-runoff model provides better discharge simulations when rainfall peaks are underestimated, as is the case for SM2RAIN-ASCAT. The better discharge simulation performance does, however, not necessarily mean that SM2RAIN-ASCAT is a more accurate precipitation dataset. In our opinion, the evaluation using EOBS is more informative in this regard.

Third, the analysis presented in the short comment includes only three precipitation datasets. We feel this is insufficient to draw robust conclusions about the differences in ranking between the rainfall and discharge assessments, particularly since SM2RAIN and CMORPH exhibit rather peculiar issues (strong underestimation of peaks and winter precipitation, respectively; Beck et al., 2017).

Fourth, regarding the RMSE metric, we argue that it should be avoided for the evaluation of precipitation datasets at (sub-)daily time scales as it can yield misleading results. This is due to the high skewness of the precipitation distribution and the prevalence of temporal mismatches between estimated and observed precipitation peaks. We will illustrate this with the following example. Take two precipitation products: (1) the original ERA-Interim (i.e., drizzly and underestimated peaks) and (2) an improved version of ERA-Interim with a perfect cumulative distribution function (CDF; i.e., same temporal dynamics as the original but less drizzly and no peak underestimation). For many locations, the improved ERA-Interim would yield a worse (i.e., higher) RMSE score because the higher peaks result in larger RMSE values when there are temporal mismatches between estimated and observed peaks, which is frequently the case. The KGE does not suffer from this problem, and would give a better (i.e., higher) score for the improved ERA-Interim. The RMSE values presented by Dr. Brocca in his short comment are likely affected by this issue and should therefore be interpreted with caution.

Secondly, the results obtained over CONUS are quite different from those we obtained here in Europe. Particularly, we want to underline the good performance of SM2RAIN-ASCAT dataset, mainly in terms of discharge assessment. This question about the representativeness of the results obtained in one region with respect to other regions. Several other comments can be raised analysing in details the figure, but they are not suited for a short comment.

As mentioned before, the global-scale evaluation by Beck et al. (2017) shows very similar results between Europe and the CONUS for SM2RAIN-CCI (which is very similar to SM2RAIN-ASCAT) and the other precipitation datasets, suggesting that the results of the current study are generalizable to Europe. Nevertheless, we recognize the need for more research, as explicitly mentioned in the last paragraph of the Conclusions.

As a final comment, we want to underline that we should be cautious in saying that the results obtained over a specific region or with a specific score can be used “as a guide to choose the most suitable precipitation dataset for a particular application.” We believe that more research is still needed and a significant effort linking satellite, meteorological and hydrological community is needed for a robust assessment of the precipitation datasets for hydrological applications.

Thanks for the comment. We agree and have softened the statement as follows: *“Our findings provide some guidance to decide which P dataset should be used for a particular application.”*

References

Brocca, L., Crow, W.T., Ciabatta, L., Massari, C., de Rosnay, P., Enenkel, M. Hahn, S., Amarnath, G., Camici, S., Tarpanelli, A., Wagner, W. (2017). A review of the applications of ASCAT soil moisture products. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(5), 2285-2306, doi:10.1109/JSTARS.2017.2651140.

Camici, S., Ciabatta, L., Massari, C., Brocca, L. (2018). How reliable are satellite precipitation estimates for driving hydrological models: a verification study over the Mediterranean area. Journal of Hydrology, 563, 950-961, doi:10.1016/j.jhydrol.2018.06.067.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Research: Atmospheres, 113(D20), doi:10.1029/2008JD010201.