“Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors” by Beck et al.

This is my first review of the manuscript “Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors”.

The study is very interesting and fits well with the scope of the HESS journal. It is well written and structured with relevant research questions answered in details in the results section. The literature cited is updated and figures and tables well formatted.

Despite this I have different **MAJOR** comments the authors should seriously consider:

1. 826 sensors is quite a large number for soil moisture stations and gives the impression that this evaluation is very general. However, by looking at the locations where these sensors are located the reader realizes that the majority are located over US and Europe, that is, over very data rich regions (i.e., where models tend to perform better). I think in title is much more important to highlight where the analysis is carried out rather than the number of sensors used. This give also a clearer picture of the results obtained in the study.

2. Following point 1 the results can be a bit biased towards models (also considering the type of evaluation the authors chose, see my comment 3e) and product that require use calibration (e.g., HBV runs). The product evaluation is in practice carried out exactly where in situ observations are more dense and where are more dense more calibration stations are present. This is partly highlighted by the authors but only at the end of the document while I would add more discussion about this issue.

3. The overall methodology needs to be strongly improved and detailed as many aspects are not clear and/or not well discussed and justified:

    a. The evaluation is carried by considering the temporal dynamic which is fine for the considerations done in the paper and from previous literature (see Koster et al. 2009), however, it is not clear how the evaluation at 3 hour resolution is done for satellite data with a revisit time larger than 1 day (e.g. SMAP, SMOS) and for model forced with rainfall with daily resolution. This must be clarified.

    b. The Triple Collocation (TC) is a foundation of one integration technique (i.e., the one of MeMo) and is a well known technique for the readers this manuscript point to. I am surprised that its theoretical foundation has not introduced in a more rigorous way and the assumptions made not tested. For instance, line 24 page 4 reads “with $ASCAT_{SWI}$ and HBV-MSWEP, which are independent from each other and from the passive products”. First the requirements are not the independence among the products but independence of their errors as well as their mutual linearity. These assumptions might not hold even for the products chosen (Gruber et al. 2016) as here, in addition, also SWI is systematically applied to at least two products of the triplet. This can falsify the results obtained via TC. I think some additional discussion and testing of the validity of the assumption is needed. The authors can consider the application of the Quadruple collocation technique (Gruber et al. 2016) for testing this assumption which many authors of this manuscript are familiar with.

    c. Maybe this is just a technical matter and nothing major but talking about the climatology of SMAP sounds quite weird with only four four/ five years of observations. From an evaluation point of view I think it still fine, however, from a longer perspective this climatology is likely to not consider the real climate variability.

    d. Many terms and procedures are just mentioned without specifying important details. This makes the study hardly reproducible. Examples: line 17 pag. 5 “*Temperature*

estimates were taken from ERA5, downscaled to 0.1 and bias-corrected on a monthly basis through an additive approach". How the downscaling and the bias correction has been done exactly? "Additionally, we calculated Pearson correlation coefficients for the low- and high-frequency fluctuations of the 3-hourly time series…". Please tell what correlations of high and low fluctuations would provide in addition to classical correlation?

e. This is an important aspect: "*We did not average sites with multiple sensors to avoid potentially introducing discontinuities in the time series.*" Line 31 pag. 6. This means that if the satellite footprint of a specific product includes multiple in situ stations multiple correlations values are considered? If so, this makes the process of evaluation very random and not really under control as different products are characterized by a different spatial sampling and might include a different number of stations. Moreover, this exacerbates the problem of biased results towards model or products working well over US as many correlation values would originates from stations located in United States with an additional penalization of other locations which have already less stations. For a fair evaluation each pixel must count one correlation value. In this respect the product collocation is a crucial aspect that has not properly discussed and described in the manuscript. For example in Su et al. (2015) and Massari et al. (2017) the co-location of the satellite data and model data was determined by nearest-neighbour association and a screening step for removing ground sensors non-representative at the coarse scale was implemented. In their study, if multiple valid stations co-located in a satellite pixel were present, the station with the highest mean correlation was retained (see section 2.6 of Su et al. 2015 for further details).

f. "*We calibrated the 7 relevant parameters of HBV using in situ soil moisture measurements between 2010 and 2019 from 177 independent sensors from the International Soil Moisture Network (ISMN) archive that were not used for performance assessment (Section 2.5; Supplement Fig. S2).*" Line 20 pag. 5. How the selection of these stations was carried out? Why 177? Why such a spatial distribution? Does a different choice provide similar results? I think all these aspects need to be clarified.

g. "*T was set to 5 days for all products, as the performance did not change markedly using different values, as also reported in previous studies".* The application of the exponential filter with a constant parameter T=5 days might be not appropriate for all the satellite products as the different products have a different vertical support. Since the calibration was carried out for the model why T was not calibrated also for the satellite products?

4. *"As forcing, we used the MSWEP precipitation dataset because of its favourable performance in numerous evaluations …. The calibrated parameter set was used for all HBV runs, including those forced with ERA5 or IMERG precipitation."*

I think proceeding in this way is not fair for the cross-validation. As HBV is basically a conceptual model, its parameters tend to correct also for errors contained in the data used to force it. Indeed, it has been largely demonstrated in the scientific literature (e.g., Zeng et al., 2018) that the impact of imperfect precipitation estimates on model efficiency can be reduced to some extent through the adjustment of model parameters. In other words, If you calibrate the parameters for MSWEP rainfall, then, when you force HBV with others precipitation inputs the results might be sub-optimal. Thus for a fair evaluation different sets of parameters should be used each one referring to the specific rainfall product used to force the hydrological model.

5. MeMo integration. The study is based on similar conceptual framework presented in Kim et al. 2018 here (maximization of correlation) with the difference that in Kim et al. correlations are calculated with a benchmark while here are obtained from TC. Beside the satisfaction of the underlying assumptions related to TC which I have discussed on point 3b, Eq. 3-5 of the study of Kim et al. demonstrates that for the maximization of R when merging two products (but this holds for multiple products also), cross-correlation terms must be taken into account (it is also demonstrated in Gruber et al. 2017 already cited in the manuscript) thus the framework described in MeMo integration is not theoretically optimal. However, if the products are

independent the framework collapses into a simple weighing average as cross-correlation are zero. I assume the authors consider null cross correlations within SMAP, SMOS and AMSR2 which I think is statistically not demonstrated. So i strongly suggest to provide some additional details and justifications about the integration framework used. This can explain why MeMo " *MeMo performed only marginally better in terms of R than the best-performing single-sensor product SMAPL3ESWI*" (Line 15 pag. 12).

6. *"The satellite products provided the least reliable soil moisture estimates and exhibited the largest regional performance differences on average, whereas the models with satellite data assimilation provided the most reliable soil moisture estimates and exhibited the smallest regional performance differences on average.".* I think the authors should highlight again here that this result is expected given the high density gauge observations used in the study area. Highlighting this is very important as for instance ground validation conducted in data-rich areas does not adequately reflect the added values of satellite observations (Dong et al. 2019).

Minor comments:

Line 24 pag. 3. Every satellite product contains proper quality flags for removing these low quality data while doing this with an external dataset might not guarantee optimal results. Please at least discuss this.

Line 12 pag. 4. "Three-hourly soil moisture time series of AMSR2SWI, SMAPL3ESWI, SMOSSWI". No clear how these time series are created or extracted from products having revisit times larger than 1 day. This is unknown in the paper.

Line 20 pag. 6. So the triplet is the same as above except for the presence of SMAPL3E in place of SMAPL3ESWI?

Figure 1 caption: Stations in Europe are not really visibile (e.g., Denmark). Can you make a bit darker?

Figure 2 caption: Please explain better panels b, c and d.

Line 6-9 pag. 11. I think this is the main reason.

Line 24 pag. 12, *"First, ESA-CCISWI incorporates ASCAT, which performed less well in the present evaluation, whereas".* This cannot be a reason if the integration is "optimal" as the different parent products are weighed according to their relative performance. So the second one is more likely the reason. Please rephrase or justify with more solid arguments.

Line 3 pag. 13. "*and satellite-based GPCP V1.3 Daily Analysis (Huffman et al., 2001)*" How a daily rainfall can provide 3-hourly estimates?

Line 20 pag. 14. Please explain what is the meaning of *efficiency* here.

Line 11 pag. 16. Check this sentence, it appears out of place.

Table 3: Latency of the products. Change to a more precise value or remove. Several does not provide enough information. I think ERA5 is now available with a delay of three days.

Table 3: Spatial and temporal resolution. With such a diverse range of products I suggest to replace "temporal resolution and spatial resolution" with spatial and temporal sampling.

References

Gruber, A., Su, C. H., Crow, W. T., Zwieback, S., Dorigo, W. A., & Wagner, W. (2016). Estimating error cross-correlations in soil moisture data sets using extended collocation analysis. Journal of Geophysical Research: Atmospheres, 121(3), 1208-1219.

Zeng, Q., Chen, H., Xu, C. Y., Jie, M. X., Chen, J., Guo, S. L., & Liu, J. (2018). The effect of rain gauge density and distribution on runoff simulation using a lumped hydrological modelling approach. Journal of hydrology, 563, 106-122.

Su, C. H., Narsey, S. Y., Gruber, A., Xaver, A., Chung, D., Ryu, D., & Wagner, W. (2015). Evaluation of post-retrieval de-noising of active and passive microwave satellite soil moisture. Remote Sensing of Environment, 163, 127-139.

Massari, C., Su, C. H., Brocca, L., Sang, Y. F., Ciabatta, L., Ryu, D., & Wagner, W. (2017). Near real time de-noising of satellite-based soil moisture retrievals: An intercomparison among three different techniques. Remote Sensing of Environment, 198, 17-29.

Dong, J., Crow, W., Reichle, R., Liu, Q., Lei, F., & Cosh, M. H. (2019). A Global Assessment of Added Value in the SMAP Level 4 Soil Moisture Product Relative to Its Baseline Land Surface Model. Geophysical Research Letters, 46(12), 6604-6613.