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.

We thank Dr. Massari for his thorough assessment of our manuscript.

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.

Thanks for the suggestion. We considered replacing *“from 826 sensors”* with *“from the US, Europe, and Australia”* in the title. However, since this would make the title less concise, we did not make this change. We do, however, clearly highlight in the paper that our results may not generalize to the entire global land surface and have devoted an entire subsection (3.9) to this issue.

We do not fully agree with the generalization that models perform better over data-rich regions, as this depends on the precipitation forcing used to drive the models. Our evaluation includes six models with non-gauge-based precipitation forcings (ERA5, ERA5-Land, HBV-ERA5 with and without data assimilation, and HBV-IMERG with and without data assimilation), and the performance of these models is largely representative of data-poor regions.

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.

Thanks for the comment. We have changed several existing sentences and added the following sentence to Section 3.9: *"The calibrated models (HBV and the Catchment model underlying SMAPL4) may, however, perform slightly worse in regions with climatic and physiographic conditions dissimilar to the in situ sensors used for calibration (but probably still better than the uncalibrated models).*"

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.

We agree and have added the following text to explain this more clearly in the revised manuscript: *"For the satellite products without SWI filter, we matched the instantaneous soil moisture retrievals with coincident 3-hourly in situ measurements to compute the R values.*"

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 ASCATSWI 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.

We thank the reviewer for his comment. We do not entirely agree with the statement that *"the requirements are not the independence among the products but independence of their errors,"* because if the products are fully independent, it follows that the errors will be fully independent as well. Unless of course the reference is imperfect (which is

the case if *in situ* data are used as reference), in which case the errors reflect both the product and the reference.

We recognize that the error independence assumption and other assumptions may not be fully satisfied in our study and we have therefore added the following statement to the revised paper: "*Triple collocation-based merging techniques rely on several assumptions (linearity, stationarity, error orthogonality, and zero cross-correlation; Gruber et al., 2016) which are generally difficult to fully satisfy in practice, affecting the optimality of the merging procedure.*"

We carefully examined the Quadruple Collocation (QC) methodology presented in Gruber et al. (2016). They note that QC still requires *"zero error cross covariance between some specific data set combinations"* (Section 2.4), which means that expert judgement is still needed to determine which products have correlated errors and which don't prior to estimating the correlations between two products. Pan et al. (2015) also highlighted the need for expert pre-judgement. QC is therefore only useful to estimate the correlation after already having "assumed" that particular products are more likely to be correlated than others. In light of this, we believe QC offers limited independent insight into the TC assumptions. The developer of QC makes a similar statement in Gruber et al. (2017): *"Recently, Gruber et al. (2016) proposed an extension to TCA where the inclusion of more than three data sets in the analysis allows for — at least partly — resolving nonzero error cross-correlation structures, yet a demonstration of the robustness of the method on a global scale is still pending. Therefore, one may for practical reasons neglect error cross correlations between different active or passive data sets at the cost of non optimal SNR improvements, or make a conservative educated guess for error cross-correlation levels for data sets where they are expected.*"

The application of the SWI filter was necessary to temporally match the different satellite products, which would not have been possible using instantaneous retrievals at non-overlapping irregular times (Gruber et al., 2020). We agree, however, that the SWI filter does not need to be applied to both satellite products in the triplets, and therefore in the revised manuscript we use unfiltered ASCAT data.

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.

We agree and have replaced *"climatologies"* with *"averages."*

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?

Thanks for bringing this up; we read the manuscript again to make sure no details are missing. We added the following to explain the ERA5 correction: *"To improve the representation of mountainous regions and ameliorate potential biases, the ERA5 air temperature data were matched on a monthly climatological basis using an additive (as opposed to multiplicative) approach to the comprehensive station-based WorldClim climatology (V2; 1-km resolution; Fick and Hijmans, 2017)."*

The following sentence was added to better highlight the added value of the Pearson correlation coefficients for the low- and high-frequency fluctuations: *"Additionally, to quantify the performance of the products at different time scales, we calculated Pearson correlation coefficients for the low-frequency fluctuations (i.e., the slow variability at monthly and longer time scales; R_{lo}) and the high-frequency fluctuations (i.e., the fast variability at 3-hourly to monthly time scales; R_{hi})."*

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).

We thank the reviewer for this thoughtful comment. This issue is commonly referred to as the collocation issue (Gruber et al., 2020) and unfortunately there are no satisfactory solutions, particularly when the products have such a wide range of grid-cell and footprint sizes. After much deliberation we decided not to change the current approach for the following reasons:

1. A coarser spatial sampling should, in our opinion, be penalized (as is currently the case), since it reflects a technical limitation in the ability of the product to represent heterogeneous areas.
2. We believe that grid-cells or footprints with multiple *in situ* sensors should be assigned more weight (as is currently the case), because the presence of multiple sensors reduces the sampling uncertainty and thus leads to a more reliable performance estimate.
3. The removal of *in situ* sensors that are not representative of the coarse scale is not straightforward in our evaluation due to the substantial variety in model grid-cell and satellite footprint sizes. We are not in favor of resampling all products to a common grid as this would penalize products with a higher spatial resolution.
4. The removal of 'unrepresentative' *in situ* sensors is further confounded by the fact that the location of satellite footprints varies over time (i.e., the footprint of today's satellite overpass is not exactly the same as the footprint of the next overpass). Su et al. (2015) and Massari et al. (2017) did not have this issue as their products were all gridded.
5. Retaining only the *in situ* sensors with the best performance may paint an overly rosy picture of the products.

We would like to note that our approach has also been used by numerous other researchers (e.g., Albergel et al., 2012; Karthikayan et al., 2017; Al-Yaari et al., 2019), which thus implicitly agreed with our view. Nevertheless, we agree about the importance of highlighting that several dense measurement networks exert a strong influence on the overall results and we therefore expanded the first sentence of Section 3.9 as follows: *"The large majority (98 %) of the in situ soil moisture measurements used as*

reference in the current study were from dense monitoring networks in the USA and Europe (Fig. 1) and therefore our results will be most applicable to these regions."

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.

We have added the following to the revised manuscript: *"These sensors did not have enough measurements during the evaluation period (March 31, 2015, to September 16, 2019) and thus were available for an independent calibration exercise.*" A different selection of *in situ* sensors would have provided similar results due to the low degrees of freedom (just 7 parameters were calibrated using 177 sensors). Note that HBV has been recalibrated for ERA5 and IMERG in the revised paper.

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?

We strongly considered optimizing the time lag constant *T* for each product in the revised manuscript but in the end decided against this for two main reasons. First, we did not want to deviate too much from the original data because we want to make statements about the accuracy of the original data, not a post-processed product. Secondly, we did not want to give the satellite products an unfair advantage compared to the uncalibrated models, which would likely also benefit from the application of the SWI filter (though likely not as much).

The calibration of HBV was carried out because the model cannot be run without calibration, as it is a conceptual model with parameters that do not represent physical properties of the land surface. Note that we added the following regarding the generalization of the performance of the calibrated models to Section 3.9: *"The calibrated models (HBV and the Catchment model underlying SMAPL4) may, however, perform slightly worse in regions with climatic and physiographic conditions dissimilar to*

the in situ sensors used for calibration (but likely still better than the uncalibrated models)."

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.

Our initial reason for not recalibrating HBV for ERA5 and IMERG was that we did not expect the resulting parameters to realistically represent the transformation of precipitation to soil moisture, because ERA5 and IMERG do not incorporate any gauge data and exhibit systematic errors (in mean, occurrence, and magnitude; Beck et al., 2019a). Conversely, the calibration of MSWEP has likely resulted in parameters that relatively realistically represent the transformation of precipitation to soil moisture, since MSWEP incorporates vast amounts of daily gauge data and exhibits almost no systematic errors in the study area (Beck et al., 2019a).

However, since we agree that the recalibration of HBV for ERA5 and IMERG might potentially lead to a small performance improvement, we followed the reviewer's suggestion and carried out the recalibration. The following text was added: *"To avoid giving one of the precipitation datasets an unfair advantage, we recalibrated the model for each of the three precipitation datasets (ERA5, IMERG, and MSWEP).*" The negligible performance improvement after calibration for ERA5 and IMERG (0.00 and 0.01, respectively) probably reflects the low degrees of freedom (just 7 model parameters were calibrated using data from 177 sensors) and thus limited ability of the parameters to correct for systematic errors.

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).

We do indeed, implicitly, assume null cross-correlations among AMSR2, SMAPL3E, and SMOS. This is an assumption to all TC applications that may not be fully met, similar to the assumption of perfectly Gaussian distributions. The null cross-correlations assumption cannot be formally tested as the truth is not known. One could evaluate the correlation in deviations versus *in situ* data but of course they do not represent the truth either and they are not available everywhere, so this does not solve the issue.

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).

Thanks for the comment. Even when excluding the three models with data assimilation using gauge-corrected precipitation forcings (GLEAM, SMAPL4, HBV-MSWEP+SMAPL3E), the remaining three models with data assimilation (ERA5, HBV-ERA5+SMAPL3E, and HBV-IMERG+SMAPL3E) still provide more reliable soil moisture estimates and smaller regional performance differences on average. This conclusion is thus not simply attributable to the inclusion of gauge observations in some of the precipitation forcings.

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.

We will expand our discussion of 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.

The last paragraph of Section 2.1 explains that the SWI filter was applied on a 3-hourly basis and that *“the SWI at time t was only calculated if ≥1 retrievals were available in the interval (t−T; t] and ≥3 retrievals were available in the interval [t−3T; t−T].”* Application of the SWI filter is thus certainly possible for products with revisit times longer than 1 day.

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

This is correct.

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

Thank you for the comment. We have increased the size of the stations and completely revised the figures.

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

We have expanded the caption with a few additional details.

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

The vertical representativeness could well be the main reason, however, we believe the noise reduction is also an important reason, given the often substantial seemingly random variability between consecutive instantaneous retrievals.

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.

The reviewer is right in theory; as discussed earlier in our response, given the difficulty of satisfying all triple collocation assumptions, our merging approach is unlikely to be fully "optimal," and we did not claim it was. For this reason, the inclusion of a product of lower quality results in a performance degradation. As mentioned before, we have added the following statement to the preceding paragraph to highlight this: "*Triple collocation-based merging techniques rely on several assumptions (linearity, stationarity, error orthogonality, and zero cross-correlation; Gruber et al., 2016) which are generally difficult to fully satisfy in practice, affecting the optimality of the merging procedure.*"

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?

Good question. This is explained in Section 2.1: *"Since the evaluation was performed at a 3-hourly resolution, we downscaled the two products with a daily temporal resolution (VIC-PGF and GLEAM) to a 3-hourly resolution using nearest neighbor resampling.*" We realize that this is not ideal, but there was no other solution.

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

Thanks for the comment. By efficiency we refer to how realistically the model represents the transformation of precipitation into soil moisture. We have rephrased *"the model efficiency"* to *"the soil moisture simulation efficiency."*

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

Deleted, thanks.

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.

We have provided more precise latency values. The latency of ERA5 appears to be 6 days at this moment.

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.

Done.

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.