Review #2 rebuttal

This is my second review of the manuscript "Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors". Some of the issues that made the manuscript not really clear at the beginning have been clarified and the manuscript has been improved in this respect, however, there are still some MAJOR and MODERATE pending issues the authors should address.

In the following:

a) **my new comments and replies to authors are written in bold red**
b) **authors replies to previous comments are in** *black italic*
c) **Old comments given by myself are in bold black**

My comments are listed below:

1) **MAJOR**

   This comment is related to the clarification about the 3-hourly Pearson correlation coefficient.

   I am a bit surprised the authors decide to use 3-hourly sampling to compute temporal correlation given that 13 out of the 18 products have temporal resolution >= 1 day (the majority of the models have native resolution equal to 3 hours given that they are forced by 3-hours rainfall, however, all satellite derived products plus one model forced by GPCP and GLEAM have resolution larger or equal than one day. The problem is that this forces the authors to downscale the majority of the products to something that is far from their original resolution. Of course a 3-hourly product has its strength but this can be still highlighted in the manuscript.

   Anyway, I am still fine with this approach but I have some doubts on how the downscaling has been carried out given that no details are found in the paper. For example I report below a couple of questions:

   - **assuming one product has observations every day like GLEAM, the three hourly product results from the downscaling is a product having the same daily value for all the eight 3-hourly intervals? Is this obtained product compared with the 3-hourly in situ observations then?**

- **for satellite observations the exponential filter seems to be used as an interpolator to bring the information of satellite passes (even every three days for products like SMOS) to 3-hourly sampling. The obtained 3-hourly products are compared with 3-hourly in situ observations?**

If so, it is likely that this creates an unfair evaluation between products with temporal sampling equal to 3 hours and those having native resolution larger or equal than one day. Indeed, the interpolation (downscaling) within such long temporal windows can yield significant interpolation errors. Please provide some more information on how the downscaling has been carried out and on the impact of interpolation errors on the correlation.

2) **MODERATE**

I am still not convinced about the title. Just mentioning the number of sensors does not reflect where the validation has been carried out. However, this is my personal opinion and I leave the authors and the editor the last decision on that.

3) **MODERATE**

I give my reply to the authors below.

Old comment 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.

Reply by the authors:

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.

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).”

My reply:

- **I understand your point, however, model precipitation (ERA5 for instance) assimilates a large number of ground observations like 2-m temperature and humidity and, in US -- where most of the stations of the study are located -- also the NCEP Stage IV analysis rainfall which combines rain gauges and radars estimates (Lopez et al. 2011). Therefore, models forced by ERA5 have to be considered something not far from gauge-corrected products at least in the US and will likely to perform better with respect to what they can do within data scarce regions.**
  Moreover, the calibration can significantly help to improve the performance of conceptual models like HBV where the soil moisture station density is high.
- *“but probably still better than the uncalibrated models”*. **Please either demonstrate this statement or provide a reference, otherwise remove.**

4) **MODERATE. I still doubt about the validation exercise. I can finally accept this approach, but I provided below some of the reasons behind my doubts.**

   Old comment 3.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).

Reply of the reviewers:

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.*

   My reply:

   I partially agree as due to temporal stability issues (Vachaud, 1984) it is likely the temporal dynamic of the stations is similar (so the results in terms of correlation are potentially less affected than any other metric like bias and error).

   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.

   My reply:

   True but again this will favor calibrated runs.

2. 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.

   My reply:

I agree with this.

3. 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.

   My reply:

   Fine, but this means that the stations considered at different time steps will vary in your evaluation from time to time? Can you provide more details on this? If so this must be specified.

4. Retaining only the *in situ* sensors with the best performance may paint an overly rosy picture of the products.

   My reply:

   I do not fully agree with this. For the same temporal stability issue described above, the stations with the best performance are potentially the ones more representative of the spatial mean related to the domain of the satellite footprint. So it is not unfair to consider them but, to my opinion, it would be the best thing to do.

Vachaud, G., Passerat de Silans, A., Balabanis, P., & Vauclin, M. (1985). Temporal stability of spatially measured soil water probability density function. *Soil Science Society of America Journal*, *49*(4), 822-828.

5) MAJOR

Old comment 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.**

Reply by the authors:

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."

My reply:

Thanks, this is clearer now for products where the Exponential filter was not applied. However, where the Exponential filter has been applied please refer to my comment 1 above.

6) **MODERATE/MAJOR**

Old Comment:

"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).

My replies:

"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."

I think that downscaling satellite time series at 3-hourly resolution (from original revisit time of more than one day) by the application of the exponential filter (it does not matter whether T is 5, 7 or 3 days) already provides a strongly 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)."

R: Well, HBV is calibrated with 7 parameters on the 177 stations so I do not see limitations on doing the same calibration of the exponential filter with one single parameter (which, in its original formulation, is itself a conceptual approach to obtain root zone soil moisture). That is, the 177 calibration stations could be used to calibrate the parameter *T* which best fits observations in terms of correlation.

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)."*

R: remove the *"but likely still better than the uncalibrated models"* **as it is not demonstrated or provide a reference to validate this statement.**

7) MODERATE/MAJOR

Old comment 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).

Reply by the authors:

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.

My Reply:

All ERA5 runs contain gauge precipitation in the US where most of the stations are located, so in practice, only HBV-IMERG+SMAPL3E (which is a calibrated product) has in theory no gauge information in it.

8) MODERATE

Old comment:

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.

Reply of the authors:

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.”

My reply:

I think that ESA-CCI contains so many products and the merging procedure so complex that it is impossible to affirm that the guilty is one product rather than another one. ESA-CCI contains also SMOS which in Figure 2 is worse/equal to ASCAT but I do not feel to say the guilty is SMOS. Please revise this sentence or provide a more solid argument to state that.

9) MODERATE

Old comment:

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?

Answer by the authors:

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.

My reply:

Can you clarify it better? Do you downscale GPCP daily to 3 hourly data? So the daily value is divided by 8 to have consistent daily accumulations?

10) MODERATE

I think it is important to provide some plots of the time series for instance for one/two locations (to put at least in the supplementary information) to better visualize the impact of the downscaling procedure and the visual comparison between the products.