This paper describes the performance of various gridded soil moisture products within situ surface soil moisture measurements. There is a lot thrown into this comparison and the methodology seems solid, but I am not sure about what we learned in the end.

We thank the reviewer for their thorough assessment and helpful comments.

Briefly summarized, we evaluated the largest and most diverse selection of soil moisture products to date, to the best of our knowledge. This allowed us to gain several novel insights into the relative advantages and disadvantages of a broad range of methodologies and data sources used to estimate soil moisture, as well as into techniques to evaluate the estimates. Of course, one outcome of our comparison is quantitative information on the relative performance of different products, which will be helpful for researchers deciding which product(s) to use in their analysis. However, there are several other findings:

- a smoothing filter helps to avoid disadvantaging noisy satellite products in product evaluations;
- the new ERA5 reanalysis precipitation data provide performance close to gauge-based precipitation estimates;
- satellite products perform worse in cold climates than in warmer climates, and so do model products;
- a simple, calibrated model can outperform substantially more complex, data-intensive models;
- precipitation data quality is the main factor determining the benefit of data assimilation;
- satellite data assimilation provides greater performance improvements for models with a poor soil moisture simulation efficiency;
- model calibration can be more beneficial than satellite data assimilation into an uncalibrated model;
- satellite products tend to exhibit larger regional performance differences than models;

We believe that these findings, which are succinctly summarized in the conclusions, are of value to the general readership of HESS.

1. Why is this particular subset of products selected? It is mixing spatial (horizontal & vertical) resolutions, operational and research products, etc., and makes a fair comparison questionable. Furthermore, it is not possible to stratify the results based on this random mix of features. Please provide more justification for the evaluation setup or refocus the paper.

This is a good question. We have added the following to the revised manuscript to justify our product selection: "*We evaluated six products per category, which was sufficient to compare the performance among and within product categories and address the questions posed in the introduction. We only considered widely used products with (quasi-)global coverage, and we attempted to keep the selection of products in each category as diverse as possible. For example, we considered products based on several major satellite missions used for global soil moisture mapping (AMSR2, ASCAT, SMAP, and SMOS), models of various type and complexity (with and without calibration), different sources of precipitation data (satellites, reanalyses, gauges, and combinations thereof), and various data merging and assimilation techniques (with different inputs).*"

For example, why was SMAPL3E included and not the coarser-scale L2/3 product, why SMOS v650 and not SMOS-IC, why GLEAM, etc. There is no reason given for the chosen products, even though the various products serve different purposes and have very different characteristics (e.g. SMOS retrievals offer both SM and VOD, DA products offer much more than only surface soil moisture, just to name a few).

Please see our preceding response. There are too many candidate soil moisture products available for us to include all of them in our analysis. Some admittedly subjective selection is therefore necessary.

We appreciate that different products have different design objectives and characteristics and offer different auxiliary data, but do not see how that invalidates our evaluation.

Perhaps this paper should move its focus towards evaluating the new MeMo product and its underlying HBV modeling system, rather than shuffling that product into a general analysis that tries to vaguely address a list of too general questions for anon-representative or inconsistent subset of data products?

Thanks for the comment. A paper just about MeMo and HBV would be of interest to a much smaller part of the community than the present evaluation. The MeMo product

was included primarily to assess the effectiveness of a different merging approach compared to ESA-CCI. The HBV model products were added to (i) examine how well a simple calibrated model performs, (ii) assess the impact of different precipitation forcing datasets on the overall performance, and (iii) quantify the benefits of satellite data assimilation for different precipitation forcing datasets.

We believe the nine questions posed in the paper are pertinent to numerous researchers and end users of soil moisture data and not “too general.” We are not entirely sure why the reviewer refers to our way of addressing the questions “vague” as we provide clear, concise, and well-referenced objectives and findings.

As an aside, HBV is not underlying the MeMo product, they are completely independent products.

Random example: what is the relative performance of the single-sensor satellite products? If “all” available soil moisture products would be compared, or some meaningful features would be targeted, then we could learn something from this, but for the 4 discussed products, more than half of the answer was already given in earlier papers and the added value of the answer in this paper is minimal.

This example pertains to just one of the nine questions addressed in the paper. Earlier papers did not apply smoothing filters, stratified the results in different ways, most did not explicitly assess high- and low-frequency fluctuations, and most did not compare the performance of these single-sensor products to other types of products. As such, we believe our analysis does provide significant added value. Furthermore, even if part of our answer to this question was already given in earlier papers, we do not see it as a bad thing to replicate the findings of previous papers.

2. The 5-day filter is used to reduce noise, but has also been used to derive root-zone soil moisture in the past. Why are the results compared to surface soil moisture and not root-zone in situ measurements? Would that not be fairer?

We actually applied the 5-day filter to make the comparison with the *in situ* measurements at 5-cm depth more fair. The 5-day filter serves two purposes, to reduce noise and to deepen the vertical support of the superficial satellite observations (not to the root zone but to approximately 5 cm). Without the 5-day filter, the satellite products would perform, on average, significantly worse than the two other major product categories, and some particularly noisy satellite products would be severely disadvantaged (e.g., SMOS; see Fig. 2). If we had used *in situ* measurements of the

root zone as reference we probably would have used a filter with a longer temporal window.

3. In general, there is very little mentioning of the vertical representativity of the various products. It cannot possibly be that all products produce a consistent~5 cm surface product. For example, how deep is the HBV soil moisture store? Is it comparable in volume to the volume observed by satellite data or other model-satellite surface soil moisture products? Due to their different wavelengths, the AMSR2, ASCAT and SMOS/SMAP products must be sensitive to different vertical surface layers. Is it fair to compare them all to the same ~5-cm surface in situ measurements?

Thanks for the comment. We believe our study represents a fair comparison. The 5-day filter deepens the vertical support to make the superficial satellite observations more representative of *in situ* measurements at 5-cm depth (please see our previous response). Previous soil moisture product evaluations tended to compare soil moisture retrievals directly to *in situ* measurements at 5-cm depth, and therefore may have underestimated the 'true' skill of products. We considered optimizing the time lag constant *T* for each product but decided against this, because we wanted to make statements about the accuracy of the original data, not a post-processed product.

We have added the following text to the revised manuscript regarding the vertical support of the models: *"The vertical support is physically consistent with in situ soil moisture measurements at 5-cm depth for most models. The average depth of the soil layer (i.e., half the depth of the lower boundary) is 2.5 cm for SMAPL4, 3.5 cm for ERA5 and ERA5-Land, 5 cm for GLEAM, 8.5 cm for HBV-ERA5, 6.6 cm for HBV-IMERG, 7.3 cm for HBV-MSWEP, and 15 cm for VIC-PGF (Table 1; Supplement Table S1). The soil layers of HBV may seem too deep, especially since they represent conceptual "buckets" that can be fully filled with water, in contrast to the soil layers of the other models which additionally consist of mineral and organic matter. However, the soil layer depths of HBV were calibrated (see Section 2.3) and are thus empirically consistent with in situ measurements at 5-cm depth.*"

4. The temporal resolution is also questionable: how is it possible to do a 3-hourly evaluation for all products (p.3, L.20)? Satellites only pass over every so many days.

Thank you for the comment. We have added the following text into 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.*"

5. Please provide more information on the quality screening of the satellite data. The text only mentions screening for frozen conditions, but each product comes with its own flags that need to be applied. For example, it is mentioned that AMSR2 and SMOS are more vulnerable to RFI: how did you screen these data for RFI? Did you screen for dense vegetation, topographic complexity, etc?

We appreciate the suggestion. We will provide more information about the quality flags used for the satellite products in the revised manuscript. Thanks for the suggestion.

6. The consideration of both high and low frequency signals for the calculation of R is a good idea, but why is there no evaluation of the interannual variability, using a simple state-of-the-art anomaly R?

The “state-of-the-art anomaly R” measures both the (seasonal-scale) interannual variability and short-term deviations from the long-term mean seasonal cycle. The skill of short-term variations in the soil moisture products is assessed in our high-frequency filter.

We did not separately evaluate the (seasonal-scale) interannual variability due to the short temporal span of some of the products (less than 5 years), which precludes us from calculating reliable correlation coefficients.

7. Not understood: “only HBV and the Catchment model underlying SMAPL4 have been calibrated”. Is it fair to say that Catchment would be “calibrated” (for soil moisture,just like HBV?) in order to hardwire a single parameter (a constant)? Wouldn’t all models then ever have been ’calibrated’ to chose some hardwired parameters?

Both HBV and Catchment have been explicitly calibrated against independent *in situ* soil moisture measurements by optimizing a certain performance metric. The same may be true but has not been similarly documented for the other models included in the evaluation. The calibration procedure of HBV is described in Section 2.3, while the calibration procedure of Catchment is described in Reichle et al. (2019b).