This is a very interesting and promising paper certainly useful to document the biblio-graphical effort on soil moisture evaluation. I feel however that authors have skimmed over some essential explanations and was sometimes wondering if I had the latest version of the manuscript from HESSD (?). The bullet points format of the manuscript does not help and a lot of discussion is missing prior it can be considered for publication. I recommend major revisions, please see below an attempt to help.

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

Although very important, this kind of evaluation is by design almost never in favour of the satellite based products. It has been highlighted several time in the literature in the past decade that in data rich areas where models are highly constrained by high quality observations, their soil moisture is of better quality that the one retrieved from spatial remote sensing. As the in situ measurements sensors you are using are largely located in those data rich areas, this should be emphasize in the manuscript.

We agree that this issue affects many previous studies. We designed our study to give the satellite products a fair opportunity in two ways:

1. We included 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). The performance of these models is largely representative of data-poor areas.
2. We evaluated versions of the satellite products processed with SWI filter which generally performed substantially better (Section 3.2). Previous soil moisture product evaluations tended to compare instantaneous soil moisture retrievals directly to the *in situ* measurements, and may therefore have underestimated the 'true' skill of satellite products.

However, despite this, the satellite products still generally performed worse.

We agree with the reviewer that it is important to highlight that models with gauge-based precipitation forcings may not perform as well in data-poor areas, which we have done multiple times in the paper:

- *"It should be kept in mind, however, that these studies, including the present one, used in situ soil moisture measurements from regions with dense rain gauge networks, and hence likely overestimate model performance (Dong et al., 2019).*"

- *"In sparsely gauged areas the four models using precipitation forcings that incorporate daily gauge observations (GLEAM, HBV-MSWEP, HBV-MSWEP+SMAPL3E, and SMAPL4; Table 1) will inevitably exhibit lower performance (but not necessarily lower than the other models)."*

Page 1, Lines 9-8 : a) It gives the false impressions that data assimilation brings an improvement going from 0.69 to 0.72 while models with data assimilation do not all have open-loop counterparts (the opposite being true as well). I know it is the abstract but perhaps you should already give scores that can highlight the added value of data assimilation by considering the mean R values of their open-loop counterpart (HBV+ERA5, HBV+IMERG, HBV+MSWEP). b) I am personally not a big fan of such statement in an abstract and I am not sure it is well supported by your results particu-larly regarding the large distribution of your scores (boxplots of figures 2 & 3) and the lack of discussions on score difference significance

We appreciate the comment. Since the abstract is already quite long we won't be able to present median *R* improvement scores for each of the products with and without data assimilation. As suggested, we have deleted the statement referred to by the reviewer.

We have added probability (*p*) values (calculated using the Kruskal-Wallis test) in the manuscript to Table 1 where we compare the performance of ascending and descending overpasses of the single-sensor products. However, we follow the soil moisture product validation recommendations set out by Gruber et al. (2020) and avoid making any statement or interpretation about statistical significance or non-significance, because *"a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant."*

We did not present *p*-values for all 254 product combinations for all three performance metrics (R, R_{hi}, and R_{lo}) and did not explicitly report *p*-values when comparing the medians scores of different products, as this would significantly hamper the readability of the paper. Additionally, we carried out some experiments using the Kruskal-Wallis test on synthetic *R* distributions with properties similar to the actual *R* distributions, and found that even small differences in median *R* of just 0.02 tend to be statistically significant at the p=0.05 level, whereas greater differences of 0.03 tend to be statistically significant at the p=0.001 level. Thus, the reviewer can safely assume that differences in median *R* of ≥0.02 will be statistically significant at (at least) the p=0.05 level.

Regarding the medians of the major product categories (discussed in section 3.8), these are all significantly different at at least the $p=10^{-11}$ level.

Page 1, Line 14 (also Line 16 and true for many part of the manuscript): Are those differences significant? why didn't you provide confidence intervals? Also according to figure 2 it is ESA-CCI_SWI that has a median R value of 0.67 while ESA-CCI has a median R value of 0.56, please clarify. The notion of with/without SWI does not appear in the abstract (?).

The SWI subscript was indeed missing in the abstract and was added to ESA-CCI and the other satellite products in the revised manuscript. Additionally, we added a line about the SWI results. Thanks for the comment.

The difference in median R between ESA-CCI$_{SWI}$ and MeMo is quite large and indeed highly statistically significant ($p=10^{-7}$). As explained in the preceding response, we prefer to refrain from making statements about statistical significance or non-significance.

Page 2, Line 14 "Additionally, many had a regional (sub-continental) focus [. . .]" I would not say yours is different (?) Particularly looking at figure 1, please clarify. Also you could add a lot of recent references that had looked at very similar dataset to like to yours. You are only slightly discussing towards the end of your manuscript, please revised

There are numerous soil moisture product evaluations that focused on a single country or a small area (i.e., a sub-continental region), whereas we tried to use all available *in situ* data globally to draw the most generalizable conclusions possible. Our *in situ* data covers the entire conterminous US and thus can be considered at least "continental." We recognize, of course, that the coverage of the *in situ* sensors is far from fully global, and to we have devoted an entire subsection to discussing the generalizability of our results. Our study has already well over 200 references and we cite numerous recent studies that also use ISMN data as reference. We are not sure which recent references are missing.

Page 2, Lines 25-26 "Furthermore, several new or recently reprocessed products have not been thoroughly evaluated yet, such as ERA5 (Hersbach et al., 2020), ERA5-Land (C3S, 2019), and ESA-CCI V04.4 (Dorigo et al., 2017)." For ERA5, Li et al have used 842 qualified sites covering 25 networks (rather recent paper I must admit):

https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.6549 For ESA-CCI, Did you check the product website and documentation? https://www.esa-soilmoisture-cci.org/validation

Thanks for pointing us to the paper and website which are both very interesting. However, our paper was already finalized before Li et al. (2020) appeared online. We were aware of the online ESA-CCI evaluation, but we did not include it in the paper primarily because it has not been peer-reviewed.

Page 3, Line 1 "[. . .] from 826 sensors located primarily in the USA and Europe [. . .]" Thus as for previous studies you have mentioned the extent to which your findings can be generalized is unclear (?), please revise as this sentence could be misleading.

Agreed. We have replaced *"and thus the extent to which their findings can be generalized is unclear"* with *"potentially leading to conclusions with limited generalizability."*

Page 3, Line 5 Question on SWI appears only here, seems a bit out of the blue (?)please introduce SWI earlier not to confuse readers.

We agree, and have added the following to introduce the SWI: *"There is also still uncertainty around [...] the impact of smoothing filters such as the Soil Wetness index (SWI; Wagner et al., 1999; Albergel et al., 2008) on the performance ranking of products."*

Page 3, Line 14, section 2.1 I am wondering here if I have the correct version of the manuscript as several dataset are not presented? It is a general comment that you have to justify why you have used those 18 dataset and not others, otherwise it looks like cherry-picking. While some are state-of-the arts, others are self-made, please revised the choice and presentation of the dataset.

All products are introduced in Table 1, which we refer to in the first sentence of Section 2.1. We have added the following to justify our selection of products: *"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)."

Page 3, Line 24 I assume you have used soil temperature of the first layer of soil between 1-7cm, is so please say it. Alternatively you could have discarded in situ measurements of soil moisture when associated measurements of soil temperature (if available) was < 4 dC

We agree and have added 0–7 cm in reference to the ERA5 soil temperature estimates.

Page 4, Lines 16-17 Add references I appropriate

We would be happy to add relevant references but we are not aware of any. This is a relatively simple part of our methodology that we believe can be understood and replicated without references.

Page 5, Line 10 "The model was run twice for 2010–2019 [. . .]" Please clarify if this was done for each forcing dataset (I assume so)

Yes, the initialization was performed for each precipitation dataset. In the revised manuscript HBV is recalibrated for each precipitation dataset. We have added the following text: *"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).*"

Page 5, Line 20 "We calibrated the 7 relevant parameters of HBV [. . .]" This will have to be discuss further already if it impacts your results wrt to the land surface model based product?

We have added the following to the revised manuscript: *"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).*" Section 3.7 discusses the benefits and limitations of model calibration in detail, including implications with respect to the land surface model-based products, as suggested by the reviewer.

Page 6, section 2.5 Are they all using the same measurement methodology?

Thanks for the comment. We added the following text: *“The measurements were performed using various types of sensors, including time-domain reflectometry sensors, frequency-domain reflectometry sensors, capacitance sensors, and cosmic-ray neutron sensors, among others.*”

Page 7, figure 1 In such study this kind of global maps tend to show areas with no data more than areas with data. It is not obvious than 2 two zooms over North America and Europe add anything, perhaps you could have one figure with 3 panels, North America,Europe and Australia (?)

Agreed; we have revised this figure (as well as the other figures) as proposed by the reviewer, but with four panels instead of three (Alaska, Europe, conterminous US, and Southeastern Australia). Thanks for the suggestion.

Also I suspect here that most of the stations in the "cold" class over North America are from the SNOTEL network located in mountainous area where the retrieval of soil moisture from space is rather complex. This should be emphasise in the text at it is biasing your results.

We agree; thanks for the comment. The retrieval may indeed be more complex in cold regions, which we mention in the paper: *“the confounding influence of dense vegetation cover (de Rosnay et al., 2006; Gruhier et al., 2008; Dorigo et al., 2010), highly organic soils (Zhang et al., 2019b), and standing water (Ye et al., 2015; Du et al., 2018) on soil moisture retrievals.*” The influence of mountainous terrain on the retrievals is also mentioned in the paper: *“Most satellite products performed worse in terms of R in areas of steep terrain (Fig. 2d), consistent with previous evaluations (Paulik et al.,2014; Karthikeyan et al., 2017a; Ma et al., 2019), and attributed to the confounding effects of relief on the upwelling microwave brightness temperature observed by the radiometer (Mialon et al., 2008; Pulvirenti et al., 2011; Guo et al., 2011).*"

An additional explanation for the lower performance in cold regions (missing from our original submission) may be that the sensors are less representative of the coarse scale of the products. We therefore added the following: *“it could also be that the in situ measurements are [...] less representative of satellite footprints or model grid-cells.*”

Page 9, figure 2 I may have missed a point but I did not understand how did you obtain 3-hourly data for e.g. ASCAT, SMOS, ESA-CCI, SMAP...please revise.

This was indeed not clearly explained. We have added the following text: *"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."*

It would have been easier to have them close to one another (SWI and not SWI) on your figures but has you have several questions to answer it was probably not easy to pick up the correct order of products for those figures.

We agree that having the SWI and non-SWI products close to each other in the figure would be useful for answering the SWI-related question but less useful for the other questions addressed in the study.

Page 11, section 3.2 My personal opinion is that this is a low pass filter smoothing the time-series, nothing more

We agree with this observation; the SWI filter is in essence a low-pass filter smoothing the time series.

Page 11, section 3.3 Are you R values significant? I may have missed something here but from your figures 2 and 3 (boxplots distribution) it is difficult for me to give a clear answer to this question (while you are doing it in the abstract)

The large majority of *R* values are highly statistically significant, since an *R* value of just 0.14 tends to be needed to obtain a statistically significant correlation (at the 0.05 level) for a sample size of 200 (the minimum sample size before an *R* value is calculated in this study; see Figure 1). Our *R* values are, however, generally much higher (Fig. 2) and our sample sizes much greater, and therefore our *R* values will be much more statistically significant.

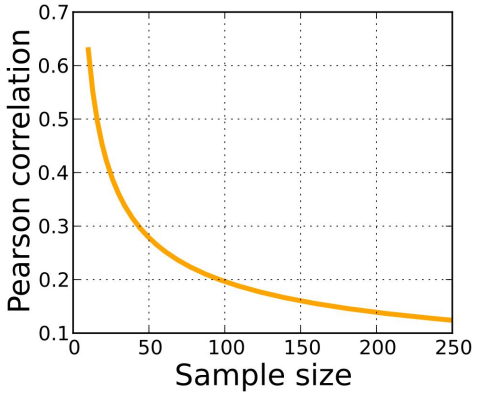

Figure 1. Plot showing the minimum value of Pearson's correlation coefficient (*R*) that would be significant at the 0.05 level for a given sample size. Source: https://commons.wikimedia.org/wiki/File:Correlation_significance.svg.

As explained in the beginning of this response letter, following recommendations of Gruber et al. (2020) we refrain from making statements about statistical significance or non-significance.

Page 12, section 3.4 Line 21 "[. . .] the central Rocky Mountains [. . .]" This are usually area where it is difficult to retrieve soil moisture form space. Memo perhaps does better than ESA-CCI but is it good? are we talking about R values going from 0.2 to 0.3 orfrom 0.6 to 0.8? From figure 4 it is difficult to see anything (at least to me). Again, are the differences significant? Lines 22-23 Confidence interval would help Line 29 Please clarify "[. . .] from the best sensor each day[. . .]"

Please see Fig. 2d of our paper for median *R* values for mountainous versus flat areas (denoted by the letters S and F, respectively) for the different products. The median *R* is 0.61 for ESA-CCI_{SWI} versus 0.73 for MeMo, which is a substantial difference (statistically significant at the $p=10^{-6}$ level).

Page 12, Lines 31-32 Is it surprising to find the 3 calibrated HBV models leading this ranking? Again I would not claim such a best to worst ranking without discussing the significance of scores.

This was somewhat surprising given the simplicity of HBV and the fact that HBV has been designed for runoff estimation in cold regions. Conversely, numerous studies have demonstrated the flexibility and effectiveness of HBV, and the model has been calibrated against *in situ* soil moisture measurements. See our discussion in the second paragraph of the subsection in question: *"This demonstrates that soil moisture estimates from complex, data-intensive models (H-TESSEL underlying ERA5 and ERA5-Land, GLEAM, and the Catchment model underlying SMAPL4) are not necessarily more accurate than those from relatively simple, calibrated models (HBV)."* Note that we have devoted an entire subsection to discussing the benefits and limitations of calibration (Section 3.7).

Page 13, figure 4 (also true for figure 5) Not sure this figure is very helpful as hardly visible (?) Perhaps you could use scatterplots, e.g. x-axis R for ESA-CCI vs in situ,y-axis R for MeMo vs in situ and then use color-codes for any classification you like.

We appreciate the comment but a scatterplot would not tell us *where* the products perform better or worse and thus would be much less informative. We could indeed use color codes to denote the locations, but we feel a map is more clear. That said, we have completely redesigned the figure and hope it is more useful now.

Page 13, Line 1 ERA5 is a coupled land atmosphere system where ASCAT has been assimilated. Could you comment on the impact it may (or may not) have when using it to force HTESSEL land surface model in ERA5-Land? is it fully independent from ASCAT?

The assimilation of ASCAT soil moisture is unlikely to have influenced the precipitation generated by ERA5, given (i) the small influence of the assimilation on the soil moisture simulations (Muñoz Sabater et al., 2019) and (ii) the vast amounts of other observations (ground and satellite) also assimilated (Hersbach et al., 2020).

P.14, Line 1 Data intensive models could also be calibrated don't you think? I personally thing it is wrong to oppose land surface model and calibrated hydrological models. Their objectives are different.

We may be misunderstanding the comment, but we do not consider it unfair to include both land surface models and calibrated hydrological models in the same evaluation. For users simply looking for the most accurate product — probably the most common type of end-user — the data source or modeling approach is not important. We fully agree that the design objectives can be different and that data-intensive models can be calibrated as well. The calibration of computationally demanding models is however more challenging, as mentioned at the end of Section 3.7.

P.14, Line 8 There is more to say from such figure as figure 5 (?) e.g. discuss the geographical patterns

Thank you for the suggestion. We have added the following: "*For HBV-IMERG, the greatest improvements were found over the central Rocky Mountains (Fig. 5), where IMERG performs relatively poorly (Beck et al., 2019a).*"

P.14, Lines 21-23 Please discuss if it is likely to be because of the inputs quality (AS-CAT/SMOS) or a methodological matter.

We explain in the sentence thereafter that it is probably a methodological issue: *"They attributed this to the adverse impact of simultaneously assimilated screen-level temperature and relative humidity observations on the soil moisture estimates.*"

P.14, Lines 26 There is also a study showing that the assimilation of ESA CCI inGLEAM leads to a decrease of quality (Brecht et al., 2018 GMD?)

We suspect the reviewer might be referring to Martens et al. (2016). However, this study shows small (not negligible) improvements in the soil moisture simulations after DA.

Page 14, Lines 32-33 Which was expected right?

This was indeed in accordance with our expectations, but this has not been explicitly discussed in previous studies (to our knowledge).

P.15, section 3.7 Perhaps this could be moved few sections above?

We appreciate the comment. However, since we compare the benefits of model calibration and data assimilation in this section, we have to discuss the data assimilation results first. It is therefore not possible to move this subsection.

P.16, Lines 16-19 In agreement with many previous studies (e.g. Albergel et al., 2010,HESS, Dorigo et a;., 2017, RSE...)

We agree and list eight previous studies that agree with our results: *"Our performance ranking of the major product categories is consistent with previous studies for the conterminous USA (Liu et al., 2011; Kumar et al., 2014; Fang et al., 2016; Dong et al., 2020), Europe (Naz et al., 2019), and the globe (Albergel et al., 2012; Tian et al., 2019; Dong et al., 2019).*"

P.17, section 3.9 Perhaps worth referencing / discussing Reichle et al,. 2019 ? Verification of the SMAP Level-4 Soil Moisture Analysis Using Rainfall Observations in Australia, https://ieeexplore.ieee.org/document/8898398

Thanks for the suggestion. We are not sure which statement of Section 3.9 is supported by the results of Reichle et al. (2019). Note that the author of that study, Rolf Reichle, is also co-author of the present study.