> This paper presents a new, extended Bayesian methodology for estimating errors of remotely sensed soil moisture. The model is inspired by triple collocation approaches (and their assumed linear error model), and in some sense, extends triple collocation to allow time-varying multiplicative and additive errors. This new methodology is then applied to show that the sensitivity of the SMAP soil moisture product is influenced by its mis-specification of the vegetation optical depth, and that this could artificially inflate estimates of vegetation, soil moisture coupling. This paper could become an important contribution to the literature – the point about SMAP is quite informative given the broad use of this dataset. Furthermore, the new error characterization technique is an important advance and could (or should) become widely used. I applaud the authors for the careful testing of the method through a simulation study and several sensitivity analyses However, as currently written, the paper is frequently lacking in sufficient detail of the methodology employed to derive its results, as I've outlined below. In particular, for each figure in the paper, what is shown in each figure and especially how it was described must be explicitly described in the text. This is not currently the case for a majority of figures. These, and a few other major concerns outlined below, need to be addressed before it can be published.

We are grateful to Alexandra Konings for the insightful review. We have added numerous clarifications, as we outline in our response.

> Major Comments:
> A) Figure 1b lists the soil moisture as an output. If I understand correctly from the text, an explicit best guess 'true' soil moisture timeseries is never determined. This is probably the conservative thing to do – I am sure the uncertainty would be quite wide. Nevertheless, some explicit discussion/warning about the fact that this Bayesian approach is primarily for determining error statistics, and that accompanying posterior true soil moisture timeseries may not be useful (or if the authors disagree with me, some justification on that, as that would obviously be very intriguing!), is warranted.

The algorithm estimates the posterior distribution of the soil moisture at each time step, as indicated in Fig. 1. What it does not yield is a single best guess, but rather a posterior distribution, although an estimate of the location (e.g. mean) could easily be derived from the posterior distribution. We now describe this in more detail in the section on MCMC sampling:

> Each sample consists of draws from the posterior distribution, or actually an approximation thereof, of all the unobserved random variables (Output in Fig. 1b). They comprise the parameter random variables (e.g. the time-dependent biases) as well the soil moisture time series, i.e. one value of \theta for each SMAP observation.

For the future, we agree that the application of this technique to product merging (i.e. estimating soil moisture by combining several products) is an interesting avenue to explore, thus building on related triple collocation results (e.g. Yilmaz, M. T., W. T.Crow, M. C.Anderson, and C.Hain (2012), An objective methodology for merging satellite- and model-based soil moisture products, Water Resour. Res., 48, W11502, doi:10.1029/2011WR011682.)

> B) Figure 2 is unclear. How is the bias defined? And how can the RMSE be greater than posterior in right-most column of Figure 2b if sigma simulation values (Table 1) are positive?

We have now defined the bias in Eq. 7, and similarly the RMSE is now defined in a separate equation (Eq. 6). The text has similarly been extended, and so has the caption.

The dot refers to the posterior standard deviation, as we now make clear in the legend. It is also mentioned in the caption and in the text.

> C) Even though the units are the same, it is a little confusing to have both the RMSE/bias and posterior on the same axes in Figure 2b, since the former represent a *difference*. I suggest splitting this into two rows. Than in the row where you show the posterior, it would also be useful to include the uncertainty of the posterior (through violin pots if necessary) and how it compares to the prior uncertainty. Is it actually much tighter, or has the mean just shifted? The bottom of page 7 mentions that "Fig. 2b shows that the posterior standard deviations are" but I only see the posterior represented by a single point.

As we state in our response to point B), we believe there has been a misunderstanding due to our insufficiently clear wording: the dot represents the posterior standard deviation. We believe this confusion arose due to the bad wording in the legend, which we have fixed. We now denote the posterior standard deviation by s_p throughout (text and figure). We contend that these quantities are directly comparable: for instance, asymptotically the posterior standard deviation of a parameter coincides with its RMSE (in a frequentist setting), provided certain regularity assumptions apply.

The posterior standard deviation is indeed considerably smaller than the prior standard deviation, i.e. the data tighten the distribution of a given parameter. For instance, the posterior standard deviation of mu shown in Fig. 2, is <0.01 m3m-3 and thus more than an order of magnitude smaller than the prior standard deviation of 0.3 m3m-3. We hope the new figure that shows the prior distributions will help readers to gauge this difference (see point O).

> D) How is Figure 3a calculated? Is this assuming perfect retrieval? It must be influenced by the type of soil (influencing the dielectric mixing model) in some way. Also, are the different lines different average levels of true tau or something else? Please mention this also in the caption and clarify the text. What happened to the tau = 0.1 line in figure b? Did you decide to no longer use it? All of these things should be explained!

We have amended the figure accordingly. We make clear that the two tau levels are the prescribed tau in the forward simulation, which is now described in much more detail:

> To compute the predicted biases in Fig. 3a), we assumed the \tau-\omega model applied and was correctly specified (temperature, dielectric mixing model [Dobson; silt loam], single-scattering albedo \omega = 0.05, etc.). For a given value of \tau_{\mathrm{true}}, we simulated the V-polarized brightness temperatures for dry and wet soil moisture conditions. These brightness temperatures were in turn the basis for estimating soil moisture by inverting the \tau-\omega model using the wrong \tau_{\mathrm{inv}} as a function of \Delta \tau. For both dry and wet soil moisture conditions, the deviation was an estimate of the retrieval bias: their mean was taken to be an estimate of the offset M, whereas their difference allowed us to estimate L. When plotted against \Delta \tau,

> M and L are increase nearly linearly and only show a weak dependence on t\tau_{\mathrm{true}}. The slope of this relation is thus well but not perfectly defined. We refer to the slopes as \mu^{\star} (for M) and \lambda^{\star} (for L), respectively. To account for the spread due to the slight curvature and dependence on \tau, we estimated the likely range of values by computing the slopes from the differences in L or M between five equally spaced values of \Delta \tau (between -0.1 and 0.1), repeated for equally many values of \tau_{\mathrm{true}} (between 0.1 and 0.6). The range of these values was \lambda^{\star}_{\mathrm{pred}} in [2.0, 3.8] and \mu^{\star}_{\mathrm{pred}} in [0.33, 0.65] m3m-3. These ranges will later be compared to data-driven estimates, thus providing a first-order assessment of the agreement between predictions and observations, despite the neglect of other retrieval errors.

Further, we state explicitly that no tau value was assumed to produce figure b. We still have kept the same colour and linestyle across the two subfigures, as the caption should make it sufficiently clear that there is no direct link between the lines.

> E) Fig 3b: The small clarification on the definition of L and M (which falls out of the model equations pretty easily) is negated by how long it takes to understand the figure because what it shows is barely described in the text. I suggest just removing this part of the figure.

The reason we included this figure in the first place is because the interpretation of L caused no small degree of puzzlement when we presented preliminary results of this study. As we suspect that some readers will skip Sec. 2 and 3, we have included this figure and we also recapitulate the meaning of the parameters in the text. While we have kept the figure, we have amended the caption in the hope that it will facilitate its interpretation. The relevant part reads

> Explanation of the bias terms, illustrated for a time-changing sensitivity L(t) and offset M(t). A varying sensitivity changes the response of the SMAP retrieval to a unit change in the true soil moisture. When it is larger than one, the SMAP data have a larger dynamic range than the true soil moisture (illustrated by the slope > 1 in the inset). The time-average value of L is l, and the temporal standard deviation of L is given by |\lambda| (length of arrow). A variable M induces non-constant offsets, and the magnitude of its temporal variability is given by |\mu|. M > 0 corresponds to a positive offset (shown in the inset).

> F) Looking at Figure 4a, it is not clear visually that L is actually more closely related to delta tau than to tau itself. Can the authors check the statistics on this (preferably at all sites)? As evidenced by the sensitivity analyses the authors needed to do, estimating tau a priori is pretty difficult. If indeed L is a better match to tau directly than to delta tau, it would be easier for the understandability of the paper, and arguably more useful for future researchers' intuition about spatio-temporal variations in SMAP baseline soil moisture sensitivity.

We believe a potential confounding by tau is an important concern, and we have included an additional scenario in Fig. 7 to address it.

This new scenario uses two explanatory variables for L and M, namely Delta tau and tau itself. As we write in the methods: "To account for a potential confounding of \tau itself, which may also have an effect on the bias estimates, we included the smoothed SMOS \tau as second explanatory variable for L and M, referred to as$\tau control." As we subsequently describe in the results, the estimates of the Delta tau lambda and mu change very little for all stations but one. This indicates that the standard inference results are not strongly influenced by confounding from this source. Also, the lambda parameter corresponding to tau is considerably smaller than that of Delta tau: medians of 0.00 and 0.16, respectively (25^{th}/75^{th} percentiles: -0.06/0.02 vs. 0.05/0.33). We hence do not believe that an additional dependence on tau, given delta tau, is a major issue here. However, we have completely revised the discussion section and now talk at length about confounding.

We have also computed estimates using only tau as explanatory variable, as suggested above, but we do not show them in the revised manuscript. The results for the tau parameters are potentially subject to confounding due to Delta tau (see above). For the South Fork site, the impression that there is a stronger relation to tau is borne out by the data to only a limited extent. The lambda parameter estimates turned out to be (10-90% posterior interval):

- standard model, i.e. only delta tau: Delta tau lambda: 0.25-0.36
- only tau: tau lambda: -0.02 - 0.11

Across all network sites, the Delta tau lambda are consistent in the sense that the posterior medians are all positive, whereas for the only tau configuration they are almost equally distributed between positive (4/7) and negative (3/7) values.

> G) More on Figure 4: The caption mentions "The magnitude of the dependence for a unit change in delta tau, lambda* is consistent with predictions by tau-omega". This is a strong-ish claim to casually throw into a caption. First of all, I'm guessing that the grey bar is some sort of model prediction from tau-omega? This needs to be explained in the caption though. It's particularly unclear since the color between the word 'model' is different than that of the grey bar. As mentioned elsewhere, the paper does not explain how it arrives at these model predictions. This has to be explained somewhere for it to be a paper that has any chance of being reproducible. Also, presumably it would not be hard to make these model predictions site-dependent (e.g. changing soil texture, estimated albedo, mean tau) – why are they constant with time? Lastly, it's unclear exactly what's going on in the right-hand column. Is it just the left hand column divided by the average delta tau at each site? If so, given that delta tau is probably as uncertain as the performance of the new methodology in this application and given that the resulting model – estimate mismatch is actually not particularly encouraging, I suggest just leaving this out. Lastly, it would be useful if there was some discussion about what the sites mean. Are the trends in lambda and mu across sites consistent with e.g. vegetation density or canopy type characteristics?

To clarify these issues, we have made several changes to the text and figure.

We have already described the extended description of the model predictions. There, we outline how we arrive at the range of model predictions displayed in the figure as well as the limitations. The reason for using time-independent model estimates is that these estimates are based on time series, i.e. multiple time instances (with changing Delta tau) are required to estimate a time-independent parameter like mu/mustar.

We have changed the colour of the word 'model' and amended the caption:

The decent model-estimation match only pertains to lambda, i.e. subfigure b), and we have revised the caption to make this crystal clear. About lambda, we write that the magnitude are "broadly consistent with predictions by the \tau-\omega model of Sec. 4.'. Conversely, "the unnormalized quantities \mu^{\star} smaller than predicted by the model."

The right-hand column shows the comparison of the un-normalized quantities to the model predictions. To make it easier for the reader to understand this panel, we now show how the un-normalized estimates are computed in a separate equation (10), and we have greatly extended the discussion of the model predictions.

Finally, we briefly discuss the apparent dependence on potential controls (like land cover). We discuss spatial patterns and the relation to land cover at much greater length in the subsection on the sparse sites. In this subsection, we write:

> There is no clear apparent dependence of $\lambda^{\star}$ on location or land cover properties; for instance, Monte Buey and Bell Ville are within < 100 km of one another, and despite the similarity in planted crops the latter's $\lambda^{\star}$ is considerably larger.

> H) Page 10, L27: I don't see why the re-analysis data error should depend significantly on delta tau at all. Why is this assumption made?

We now discuss our rationale for including the Delta tau explanatory variable in the bias model of the re-analysis data.

> The inclusion of a $\Delta \tau$-dependent bias for the reanalysis product is not driven by physical reasoning, but for statistical reasons. By controlling for the same explanatory variables for both products, the impact of potential confounders - e.g. a seasonal bias that is correlated with $\Delta \tau$ - on the bias estimates of the remotely sensed product can be reduced. If this were not done, the model would try to partially adjust the time-variable bias term of the remote sensing product to minimize the systematic differences to the re-analysis product, thus distorting these bias estimates.

> I) The baseline SMOS VOD product is known to have significant issues, because it relies heavily on an LAI-based prior (see discussion in Fernandez-Moran et al, Remote Sensing

> 2017). The SMOS-IC product has been developed specifically to get around this and early results are looking favorable. It is not yet publicly available to my knowledge, but the authors are quite willing to share. However, I am not sure SMOS VOD is the best 'true' VOD here – it will differ from the underlying ideal SMAP values due to differences in footprint, orbit, etc between the two satellites. Thus, I suggest using VOD from the dual-channel algorithm (either the O'Neill et al once currently used in the sensitivity analysis or I'd be happy to share our MT-DCA retrievals, which have somewhat less high-frequency noise and spatially variable albedo) instead of the SMOS VOD. The point in Figure 6 about the role of using optical data vs using a climatology for VOD would work just as well even without the first column in the figure.

We share the reservations with respect to the tau products. To better address them, we have made a number of changes. First, we discuss the SMAP DC results obtained over the network sites in more detail. In particular, we mention some of the issues associated with either product. Second, we now also show the SMAP DC results over the contiguous US, i.e. over the sparse sites (Fig. ?). As with the network sites, the results are very similar. Third, in response to Wade Crow's remarks, we have included a separate discussion section where we discuss errors in the tau products and their impact on interpreting the results in a descriptive and a causal framework.

Note that we continue to use the SMOS L3 product, as we hope that we are able to paint a more complete picture by showing the results obtained with two different products. Unfortunately, the other products mentioned are currently not publically available. We hope that the techniques developed in the manuscript will in the future contribute to elucidating the error structure of novel products such as the MT-DCA soil moisture.

> J) The discussion section would benefit from some more discussion about the greater implications of this new methodology. For example, this technique might work particularly well for triple collocation of land surface fluxes of water and carbon, where it is easy to imagine significant seasonality in the error terms. Do the authors agree?

This is a good point. We have included a separate discussion section, where we dwell on the implications for error characterization more generally.

> Geophysical products in general are potentially also subject to time-variable errors, so that the presented approach could be applied to variables such as wind speed, land surface fluxes and leaf area index. The issue of non-constant error sources, be they associated with environmental conditions or varying observational parameters, likely pertains to many such variables. Extensions of our approach could in the future shed light on the error properties of a wide range of products, thus contributing to the development of improved retrieval approaches.

> K) Similarly, can the authors discuss the implications of the normalization in Eq. 6 for the interpretation of the results?

We do so by comparing the normalized results with absolute (unnormalized ones: the quantities with an asterisk). We detail the associated changes to this point in our reply to point G).

Minor Comments:

L) Page 2, line 32: See also Momen et al, JGR-B 2017

We have added a reference to this paper.

M) Page 3, line 5: You haven't defined delta tau here

At the beginning of the paragraph, we now write 'We hypothesize that seasonal changes in the error structure arise due to an inaccurate vegetation correction in the retrieval, so that the biases relative to the in-situ data track the misspecification in the vegetation optical depth \Delta \tau.' This is not a precise definition, but it should suffice for the introduction.

N) Figure 2: it would be helpful to explicitly explain somewhere why there are no RMSE values in the no mu, no lambda, no kappa case. It would also be easier to read the axes if there were more horizontal tick marks in each row, and if the tick labels were repeated between part a and part b.

We have amended the caption accordingly. We have also changed the ticks and labels as suggested. Note that we have slightly redesigned the figure in line with other suggestions.

O) Section 2.1.3: You assume quite specific priors. Would be helpful to show these distributions in the supplementary material to give the reader a sense of what they look like?

Good idea, we have added a new figure (supplement).

P) Page 7: I suggest defining the RMSE error with equation or at least separate symbol for clarity. It's easy to miss this definition in the middle of the writing, but integral to following the rest of the discussion

Done.

Q) Page 7, line 29: How is this calculated?

We now state explicitly the dynamic range on which these calculations were based: "The sensitivity coefficients \lambda are retrieved with comparable precision: the RMSE of 0.05 corresponds to a differential bias between dry and wet conditions of around 0.01 m3 m-3 (assuming a soil moisture dynamic range sim 0.25 m3m-3"

R) Figure 2: Suggest splitting this into three columns: one with posterior vs. prior distribution (in violin plots if necessary), then third column with bias and RMSE.

We believe this comment is to do with our bad wording in that we referred to the posterior standard deviation as the posterior (uncertainty), see points B) and C).

While we agree in principle that showing the entire posterior distribution is a good idea, we believe the well-behaved unimodal distributions, which we have exclusively encountered in our analyses, warrant the restriction to location and dispersion parameters to summarize those posterior distributions.

S) Figure 2: Need to make it clearer that the 'no kappa' and no mu, lambda, kappa'

simulations are cases where still have that in forward model. This is very difficult to pick out from text as is.

done

T) Page 11, line 5: note that this reference is broken

We have added the year

U) Page 16, line 1 : The authors might want to cite Crow et al, GRL 2015 here, which showed this point quite convincingly for soil moisture –latent heat coupling

We are grateful for this remark, as we were not aware of the paper.

V) I don't think the subscript p is ever defined. Is this an index for the number of explanatory variables?

We now also define it explicitly ("sensitivity to the pth explanatory variable")