“Probabilistic inference of ecohydrological parameters using observations from point to satellite scales” by Maoya Bassiouni et al.

Response to Marc F. Müller (Referee #2)

The authors use soil moisture observations in a Bayesian inversion procedure to estimate vegetation-related drivers of soil moisture dynamics in the root zone, as modeled by a simple model of soil moisture distribution. The authors apply the approach to a diverse sample of study regions where soil moisture and climate observations are available at different scales. The presented research is important and innovative in that it investigates the potential for recent remote sensing approaches that monitor spatially aggregated soil moisture to estimate eco-hydrologic parameters that are very challenging to observe in-situ, even in well instrumented basins. The research also bridges the gap between different observation scales, which has potentially interesting implications in poorly gauged regions. While I recommend the paper for publication in HESS, I would also like to raise a few comments/questions that could possibly help the authors during the revision of their paper.

Thank you for your thorough review and constructive suggestions. We have provided responses and some preliminary corrections below.

Major comments

1. The authors appear to use the same sample of soil moisture obervations to calibrate (via Bayesian Inversion) and validate (KS tests and Fig 3) the approach, which instictively raised red flags on a first read. After reflecting, it became clear (well, to me at least) that the purpose of the exercise was to show that Bayesian inversion can be use to estimate vegetation-related drivers of soil moisture using soil moisture time series, conditionnal on the assumed pdf model being an accurate description of soil moisture dynamics. In that case, the research design would be appropriate because the posterior CV portrays estimation uncertainties and the goodness- of-fit shows that the soil moisture model is, indeed, appropriate. Consequently, the purpose of the goodness-of-fit test appears to be to evaluate the functional form of the pdf, not the estimated parameter values, so it is fine to use the same dataset to calibrate parameters and evaluate outcomes. Please clarify the distinct function of these two metrics as appropriate.

Yes, the above comment describes our intentions. We will revise section 2.3.2. to explicitly state the evaluation goals and metrics used.

Optimal analytical soil saturation pdfs are evaluated by the following criteria.

(1) The Bayesian inversion converges and the Gelman-Rubin diagnostic approaches 1 for each estimated parameter (<1.1).
(2) There is goodness of fit between the optimum analytical pdf derived from the mean parameter estimates and the empirical pdfs derived from observations using the Kolmogorov-Smirnov (KS) statistic and the quantile level Nash-Sutcliffe efficiency (NSE) (Müller et al., 2016).
(3) Posterior distributions of parameter estimates are physically plausible and have low coefficients of variations.

2. I am having issues with the way you use KS tests to evaluate pdf fits. First off, if I am not mistaken, the null hypothesis of a ks test is that the two tested distributions are identical. If so, the p-value could be interpreted as the probability of obtaining a ks-distance at least as large as the one that would be obtained if the two samples were taken from the same distribution. This is loosely equivalent to the probability of falsely rejecting the null. In other words, a p-value of 5% would mean that one has a 95% chance of being right when stating that the two distributions are different, which is quite a low standard when assessing goodness of fit. Significance levels don't tell anything about type II errors, which is what I would think we are ultimately after when evaluating goodness of fits. More importantly, the KS satistic does not follow the kolmogorov distribution (i.e. estimated p-values are wrong) if the same sample of data is used to calibrate the cdf model and construct the empirical cdf to which it is compared. In my opinion, however, a formal test is not necessary to make your point here (see point 1). The graphs in Fig 3 are sufficient to make the point that the laio model reproduces the shape of the observed empirical histogram. You can then use a distance measure to monitor fits in the sensitivity analyisis. The KS-distance is probably not the most appropriate measure for that though, as it only considers the largest distance between the cdfs âA Ť global distance metrics like the Cramer Van Mises statistic or quantile-level nash sutcliffe efficiency, Muller 2016), or information based criteria (e.g, AIC, Ceola 2010) are useful alternatives to consider.

We agree that the KS test has disadvantages. The KS test was the most strict in quantifying divergence between the analytical and empirical pdfs. In contrast, the NSE values were almost always greater than 0.95

and were less useful. We agree that the p-value for the KS test is not always meaningful and will remove p-value details in the plots and text. We will report both KS and NSE values in the revision.

3. Your sensitivity analysis on soil depth (Section 4.2.) convinces me that the value assumed for Z in eqn 2 has little effect on the modeled soil moisture dynamics. This is of course important, but without actually measuring whole column soil moisture, I fail to see how you test the homogeneity assumption (i.e. that near surface soil moisture observations can be used to estimate whole-column characteristics). Please elaborate.

We agree that it is difficult to test the homogeneity assumption through the sensitivity tests in this analysis. We have decided to remove the sensitivity analysis related to Z. We will only consider Z equal to the actual measurement depth for each sensor in the revision.

4. I would find it interesting to elaborate on the interpretation of convergence in the context of Bayesian inversion. You mention (I think) that MCMC runs do not converge if insufficient information is available in the empirical p(s) to determine the considered model parameters. I would find it interesting to elaborate on when (and why) these non converging runs arise, perhaps in your discussion on data availability (section 4.3).

We agree that this is an interesting aspect of this study and will amend the results and discussion section to elaborate on the interpretation of convergences. The convergence and inference uncertainty obtained through the Bayesian approach provides insight on (1) whether the data is consistent with the model form used: whether the model is not complex enough or too complex and equifinality arises and (2) whether the assumptions necessary in the model are met by the data: whether the data spans an appropriate range of values and whether the data meets the stationarity assumption.

5. Finally, I would find it useful for get a sense of how parameters estimated using SM observations taken at a certain scale are valid at different scales. This would have interesting implications, for instance in terms of using satellite remote sensing SM observations to estimate smaller scale SM dynamics in ungauged regions. You discuss this point a little in the paper, but it would be interesting to substantiate your arguments with some analysis. For instance you could run a goodness of fit analysis between modeled SM distributions using params estimated at one scale to empirical SM pdfs observed at another scale.

These are interesting questions that may be better answered with a different dataset. We will add a few sentences describing the potential of the proposed methods to address theses questions in the discussion section and relate to recent references on scaling of the stochastic soil water balance model. Our results indicate that the parameters estimated at one scale are not applicable at other scales. One reason is that soil texture constraints (s_h and s_{fc}) are different. Another point is that when averaging over larger areas, the effects of a large number of plants (as opposed to one in a point measurement) will change the s^* and s_w. Ideally soil water retention parameters would be accurately known and soil saturation thresholds could be converted to more universal values such as soil water potentials and therefore be more transferable for scaling analysis, assuming E_{max} is uniform within the area.

Minor comments

p3. I would find it useful if you could comment on the advantages of using the Bayesian inversion approach you propose vs more "standard" frequentist approaches such as maximal likelihood, which is the go-to approach I would take to fit a "low dimensional" (4 params) closed form analytical pdf.

This comment will be addressed by revising the following sentence in the introduction:

We selected a Bayesian inversion approach instead of a maximum likelihood approach because it quantifies the inference uncertainty directly and improves upon the work of Miller et al. (2007), which used a least-squares approach to calibrate soil saturation pdfs. In addition, inference uncertainty provided by the Bayesian approach can be used to evaluate the validity assumptions necessary for the model inversion.

p7 l.18. To illustrate your claim, it would be useful if you could present statistics on the frequency of s in each zone of the pdf (in eqn 2) using your best estimation of s and sw at each site.*

We will visualize s* and sw in Figure 3 to address this suggestion. Also, we will report minimum and maximum observed soil saturation for each site and scale in Table 1.

p7. Please describe your procedure to compute empirical pdf's from time series observation. If you use kernels to estimate density functions, please specify and justify the chosen shape and bandwith.

Empirical pdfs were visualized with histograms in Figure 3 using 20 bins, evenly spaced between 0 and 1. In the Bayesian inversion, for each observed soil saturation value, we compute the theoretical probability of that value given a set of model parameters. To compute the quantile level NSE, we compare the quantile score of the observations to the theoretical quantile score from the optimal analytical pdf model. To compute

the KS we compare the *n* observed saturation values with *n* randomly sampled saturation values from the optimal analytical pdf model.

p9 l.20: 'discarded'

This will be corrected

p10: section 3 is missing

This will be corrected

p11. It would be useful to summarize the results (model, scale, posterior CV, goodness of fit distance) for the different cases of the sensitivity analysis in a table.

We will consider reporting the most important summary statistics in a results table if these cannot be clearly reported in the text and are not are already visualized in the figures.

p12 l.27. "Consistent" has a very specific statistical meaning (asymptotically unbiased), please rephrase if necessary

We will rephrased to:

when the mean and standard deviation of the randomly selected observations were most representative of the full record and therefore consistent with the rainfall characteristics**.**

p13 l 18: "versus"

This will be corrected

p 14 l3. Please elaborate on how you could disentangle confounding effects of scale and observation depths. The way I understand it, your analysis in Section 4.2 shows that the results are insensitive to the assumed root-zone depth, not the actual depth, which appears to be unknown (see point 3 above).

Yes, our analysis shows that estimates of s_w and s* are not very sensitive to the depth assumed in the model inversion. This is important if the sensing depth is not precisely known or is variable in time and space, which is the case for the cosmos and satellite measurements. We will remove the sensitivity test related to soil depth because it is not useful to determine whether estimates of s_w and s* derived from surface soil moisture measurements are relevant to deeper soil depths. We will explain this choice in the revised methods section.

Fig 5: you state that the Kolmogorov statistic is significant with a 95% confidence levels. Does that mean that the statistic is significantly different from zero? If so, I would interpret that as having a 5% chance of being wrong if I state that the two compared distributions are different (see my point on KS tests above), which I don't think is the point you intended to make.

Yes, that was the point we intended to make, we have decided to remove the details about the KS significance in the figures as response to your comment above.

References

Ceola, Serena, et al. "Comparative study of ecohydrological streamflow probability distributions." Water Resources Research 46.9 (2010).

Müller, M. F., and S. E. Thompson. "Comparing statistical and process-based flow duration curve models in ungauged basins and changing rain regimes." Hydrology and Earth System Sciences 20.2 (2016): 669-683.