We want to thank Anonymous Referee #1 (AR1) for her/his supporting and motivating words as well as for the valuable and positive criticism. First, we accepted all the minor suggestions, all of them will be implemented.

Below are our responses to each of the main comments.

In addition, the authors make often statements which are hard to find back in the figures or need a bit more support. For example, based on Figure 8, the authors state that there is a satisfactory performance, but Figure 8b seems quite off in my view, so I am not sure if you can make this statement. More importantly, there are many more tensiometers and events, so why not show more (ideally even all) data? This would generalize and support the statements of the authors much more. The same applies to Figure 11, even though the correlation coefficients are reported in Tables 7 and 8, more results can easily be shown. In addition, a more critical look at this figure is probably justified, instead of just looking at the R-squared values. For example, why are the soil moisture values flattened off? And why is the SWI just changing after a soil moisture value of around 0.155 cm3 / cm3? Just looking at the Rsquared value gives a good indication of the linear relation, but it does not say much about, for example, a consistent bias, so I might be good to look a bit further here. Related to this, I also think it is important that the authors assess the spatial correlation between SWI, HAND and TWI a bit more thoroughly, instead of just a linear regression. For example, normalizing them and subtracting the maps from each other will lead to insights where the values match, and where deviations start to occur. This can also easily be done for the interpolated values of soil moisture (Figure 13)

We agree that more results can be presented to support the claims made by us in the article. We can add graphs with all the data used in the analysis. We can also analyze the spatial correlation between the SWI, HAND and TWI indices in more detail, instead of linear regression, as suggested by the reviewer.

Regarding the calibration, what is actually the rational behind using the DREAM algorithm? First, I am a bit confused on how it was implemented, there is a mention of a generalized likelihood function, but in the next sentence (P9.L174) it is stated that the last 7500 samples are used to represent uncertainty. So how was the uncertainty actually represented in the end? In addition, the DREAM method is applied to each event to get parameter estimates, but in a next step, the parameter values of the different events are just averaged, which seems a bit simplistic in contrast to the DREAM algorithm. Why not give more weight to the parameters that are more likely? The added value of this method is also that it gives you uncertainty values, but afterwards, the authors do not really do anything with it in the analyzes of soil moisture indexes. So what is the point of using this method? In addition, I think there is also a strong influence of the chosen evaluation statistics, as the Nash-Sutcliffe efficiency and the RMSE have a strong bias for large values. In other words, in the event based approach in this study, when the height of the peak matches, high values for these metrics are likely to be found. The authors also never specifically mention how they deal with the initial states, which will have a strong influence on the results. Is it correct that these are calibrated?

We agree that the uncertainty analysis can be better detailed in the sections of Methods and Results. In this work we use a generalized likelihood function proposed by Schoups and Vrugt (2010), which relaxes the commonly assumed premises on residual errors. We do not detail the results regarding the parameter uncertainty, nor the analysis of residual errors. This occurred because initially we did not want to take the focus away from the main analysis, which is the presentation of the SWI index as a parameter representative of the variability of soil moisture over time. In fact, given the difficulty in representing the real uncertainty inherent in the calculation of the SWI, we consider removing the uncertainty analisys from the work, and simply use a standard calibration method as a genetic algorithm; we can even do this if the reviewers find it more consistent. But we can include one more item in the results section showing the uncertainty of the model parameters in the flow generation. As for the initial conditions of the model, these are considered parameters (S1I and S2I), and were obtained through calibration.

Lastly, I also wonder if it is really a surprise that TWI, HAND and SWI show good correlations. In the end, they are all based on the flow direction map, and especially on the event time-scale for a small basin, I think they should show similar patterns. I am not sure about this, but I just wonder what the authors thoughts are here. Would it not be much more interesting to see if these findings still hold for a longer time-scale and / or even a larger area?

The three indices are actually based on a map of flow directions, showing similar patterns for this studied basin. The SWI has the particularity of representing the variability of these patterns over time, and we believe that demonstrating this, even for a small basin, is a promising result. Verifying whether this standard remains in a longer time-scale or a larger area would require field data that we unfortunately do not have.