We would like to thank the Anonymous Referee #1 for this review and the constructive comments. We would like to take the opportunity to address the four major concerns immediately and address the minor comments in a more detailed response later.

General comment

The paper tests the capability of simple meteorological drought indices to detect drought events, as defined by simulated soil moisture time-series. The topic is of interest for practical applications in drought monitoring, since simulations are often hard to be performed over some areas.

My opinion is that the overall quality of the paper is negatively affected by some basic assumption made by the authors during the analysis, which are not clearly presented and sometime poorly described. Often the reported results seem off, due to errors or unclear explanations. Hence, I suggest to the authors to carefully reread the paper before to proceed with a full evaluation of the paper.

Response: All the authors are confident that the paper is ready for review. We regret and partly disagree fundamentally with some of the referee's views. We hope we can clarify the referee's major misunderstandings – see our responses below - and we await with interest the comments of other reviewers.

Specific comments

Comment 1.1: First of all, they compared the 3-month SPI and RDI against a time-series of monthly minimum pF. If I have understood correctly, this means that this time series is obtained by choosing the minimum pF value (out of roughly 30 values) for each month in the simulation period. If this is the case, I'm really surprised to see the really good correspondence between SPI and pF as shown in Fig. 4 (and 5 as well). Since SPI (as well as RDI) is a standardized variables, its "random" behavior in Fig. 4 is justified, but the same cannot be said for minimum pF which should retain a sort of seasonality depending on the climate of the area. I'm not familiar with the climate of the specific study region, so it is possible that this behavior is due to the peculiar climate of the region, but in general it is advisable to perform a correlation analysis between a standardized variable (SPI) and a non-standardized one.

Response 1.1: SPI and RDI are not standardised seasonally in this application (Eqs. 1 and 2) and therefore they do include seasonal patterns. In this regard the good correspondence between the SPI and pF should be no surprise.

In Figure 4 we plotted the simulated soil water pressure for Bourke, where the ratio between winter and summer rainfall is 0.61 (Table 1). The seasonality patterns are much more distinct Cairns though (see revised Fig. 4), where the ratio is 0.10.

We strongly disagree that "in general it is not advisable to perform a correlation analysis between a standardized variable (SPI) and a non-standardized one". The definition of any index is rather arbitrary and usually specific to the pre-defined problem. That said, we believe it is essential to compare indices with physically measurable and plausible variables – no matter if the index includes a standardisation process or not. Ideally, indices such as the SPI or RDI are compared with empirical field data. However, for a variety of reasons such long-term monitoring programs are restricted to limited funding and time. Therefore, a logical step before implementing any long-term campaigns is to test their feasibility in a desktop study using physically based models such as Hydrus-1D with available empirical data such as rainfall/evaporation and soil water retention characteristics. We believe our study covers exactly this critical step!

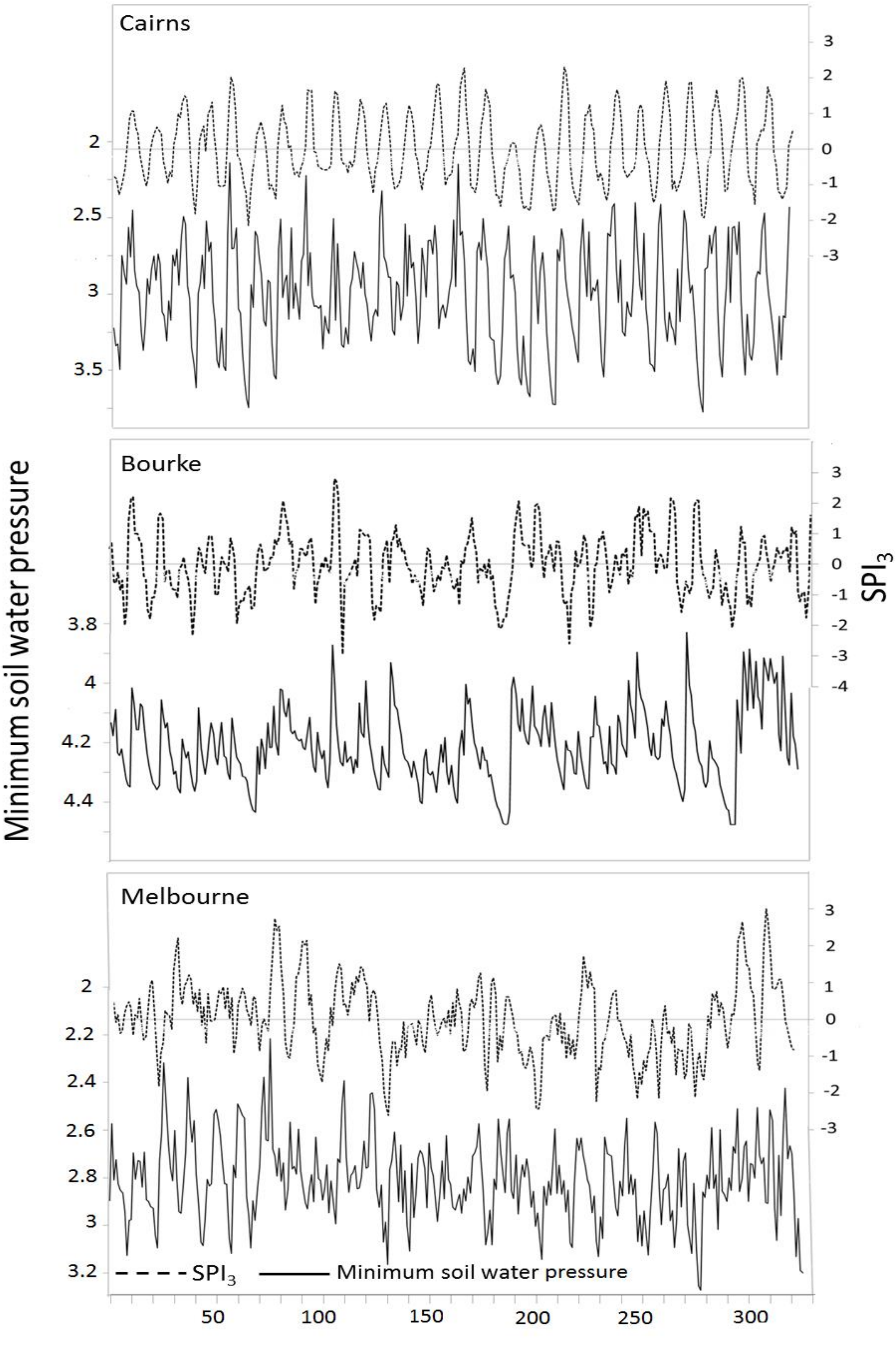

Figure 4. Simulated monthly minimum soil water pressure over 5 cm depth and SPI for Cairns, Bourke and Melbourne.

Comment 2: Also, the authors do not clarify if the 75% threshold is computed separately for each month or for the whole dataset. I assume is the first case (based on the data in Fig. B), but this is never clearly stated. Following this topic, in the same figure it seems that the 75% threshold corresponds to an SPI value around 1.2. This means either that: 1) both tails of the distribution are accounted in this computation, but the correct approach would be to consider just one tail since drought event (i.e., extreme dry conditions) are analyzed here, or 2) the fitting of you distribution is poor since the theory suggests that only about 11% of the data should be < -1.2 according to the normal distribution (about 3 values). It is fundamental that this issue is clarified and eventually fixed.

Response 2: The assumption that the *75% threshold is computed separately for each month* is wrong. In Section 2.3 (P5 L 17-18) we state that "For each site the threshold that determines a soil moisture drought event was selected by the percentile of **all simulated** pF^5 and pF^{30} …"

Further the referee has assumed that the *75% threshold is computed separately for each month* based on Figure B. We regret that the conceptual schematic in Fig. B was considered as empirical data and will clarify that in a revised version of the manuscript. In this regard, the further comments of the referee under comment 3 are irrelevant.

Comment 3.1: The analysis on extreme values is really misleading, and it also needs to be extended by including other indices. The authors say that FR and FAR are identical in all the cases, but this shouldn't be the case. FR is equal to FAR only if a and c are the same, but this is really unlikely to happen in real cases. For instance, in your example in Fig. B (which I assume is from one of your cases): FR = 5/8=62.5% whereas FAR = 4/7 = 57.1%. Please recheck your calculation of those indices.

Response 3.1: The RDI was excluded from the analysis of extreme values because of the better performance of the SPI performed better (Table 3). The RDI results may easily be added to the Appendices though upon further reviewers' comments.

We regret that the reviewer has been misled by the unequal number of data points in $a+b$ and $b+c$, which will be addressed in a revised manuscript.

Comment 3.2: Also, FR and FAR are not the only indices relevant in this case, e.g., what about the skill of the SPI? Is it better than the climatology or the random case?

Response 3.2: It is not clear to us what the reviewer suggests by the 'skill' of the SPI. It will be easy to add additional performance indices, although our current view is that the visual assessment, R2 values and FR/FAR are sufficient.

Comments 4.1: Finally, the results of the sensitivity analysis are surprising and need some clarifications. In almost all the case you have FR/FAR values between 30 and 50% higher than in the case of SPI. This means that FR/FAR values for the perturbed simulation are in the order of 65-70% in all cases, included several cases where only a 10% error in 1 parameter is added/subtracted (ie., Bourke 5 cm, Cairns 5 cm, Melbourne 5 cm). I'm really surprised by this result, since in my experience, even for a very sensitive parameter, a 10% change can rarely leads to have 2/3 of the previously detected extremes not detected anymore.

Response 4.1: The Richards' equation is used in the Hydrus model (P4 L24). Given the non-linearity of the water retention curve it is not surprising that even ±10% changes in the van Genuchten parameters affect the previously detected values disproportionately (Šimunek et al., 2012).

Comment 4.2: It would be useful to have a figure with the reference and perturbed simulations (only the maximum and median ones), as well as the corresponding threshold values, in order to better understand how these changes affect the results.

Response 4.2: We agree with the referee's suggestion, however we are concerned about the number of graphs (3 sites, 2 depths, 3 model parameters = 18 graphs). Therefore as an example we propose one graph for Appendix D – see Figure D.

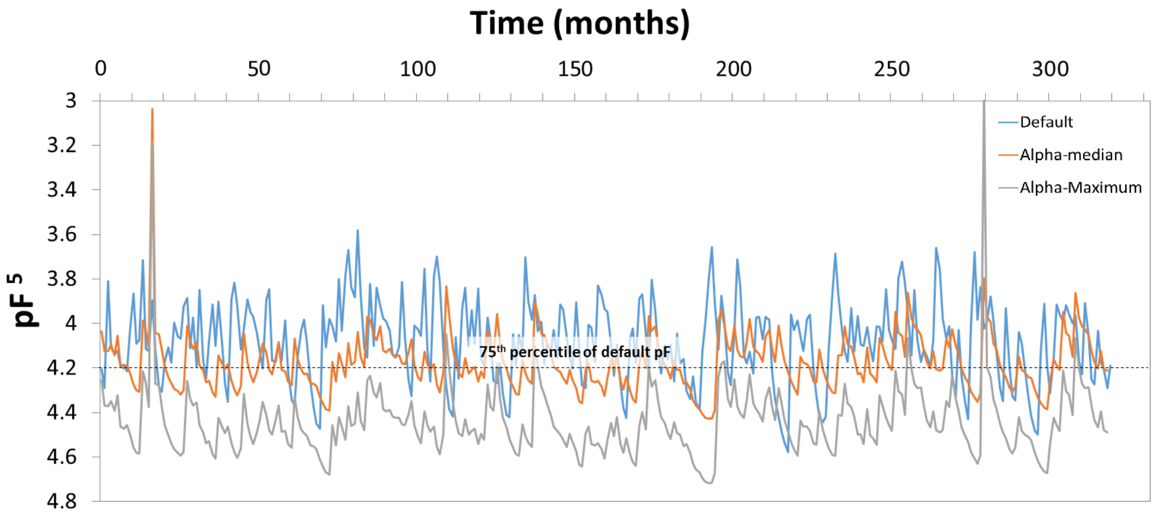

Figure D. Default and perturbed (median and maximums for parameter alpha) monthly minimum soil water pressure over 5 cm depth for Bourke.

Comment 4.3: Also, judging from Fig. B it seems that the same 75% threshold is used for both the reference and the perturbed simulation. I assumes that this is not the case, and it is just a coincidence, but I suggest to clarify in the text that the 75% threshold is adapted for each simulation accordingly to the simulated values.

Response 4.3: The reviewer's assumption is incorrect. The threshold of the perturbed pF is same as the 75th percentile of the default pF, as shown in Figure B and stated in the caption. It would not be a useful performance analysis otherwise. We will emphasise this in the revised manuscript.

Reference

Šimunek, J., Van Genuchten, M.T., Šejna, M., 2012. HYDRUS: Model use, calibration, and validation. Transactions of the ASABE 55(4) 1263-1274.