We would like to thank the Anonymous Referee #1 for this review and the constructive comments. We have addressed the referee's major comments (addressed earlier in the discussion stage) and the minor comments are as follows.

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 that the addition of new text to the manuscript will address some of the issues associated with the analysis, which were not clear to the reviewer and/or required further information. We await with interest the comments of other reviewers.

<u>Specific comments</u>

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 is expected.

The correspondence and seasonality of the SPI and pF can be seen in the raw data plotted in figure 1. Originally we only plotted the simulated soil water pressure for Bourke, which has limited seasonality - where the ratio between winter and summer rainfall is 0.61 (Table 1 in manuscript). We have revised this figure to now include Melbourne and Cairns, where Cairns shows much more distinct seasonal patterns (with a ratio of 0.10) so that the seasonality of the locations and the lack of standardisation is clearer (see revised Fig. 1 below).

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, such empirical data are often not available (such as in our study locations) for a variety of reasons, though primarily because long-term monitoring programs are restricted due to limited funding and time. The lack of empirical data is an issue across the world especially in developing nations or nations such as Australia with little history of long-term monitoring programs. Therefore, a logical step in absence of such data is to apply physically based models such as Hydrus-1D with available empirical data such as rainfall/evaporation and soil water retention characteristics. We believe our study addresses this critical step!

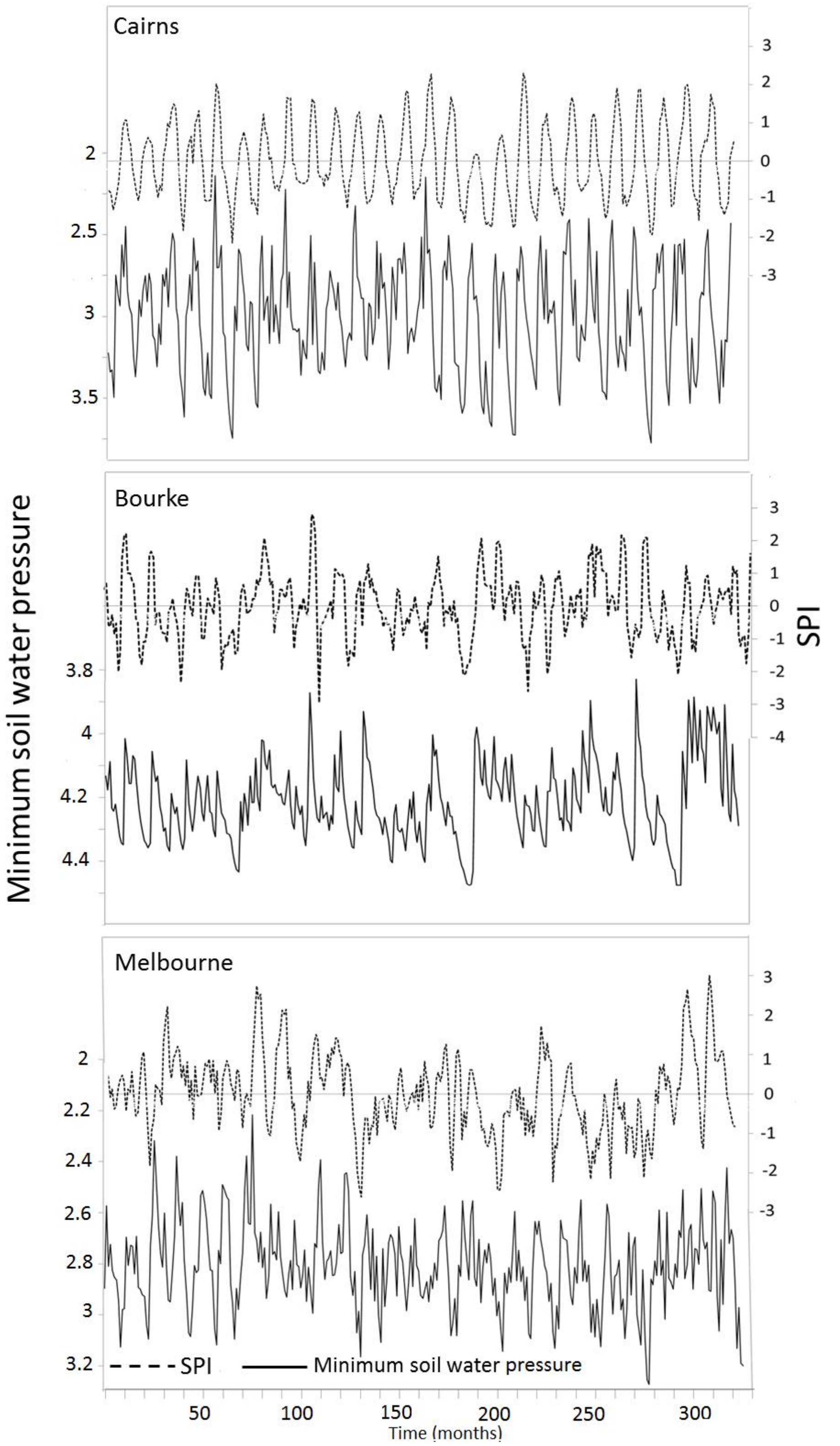

Figure 1. Simulated monthly minimum soil water pressure in 5 cm depth and SPI for Cairns, Bourke and Melbourne.

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 incorrect. 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 may have caused a misunderstanding as what was shown in the diagram was example data. We 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 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 Fig B (a+b and b+c). This will be addressed in a revised manuscript (see revised App. B (Fig. 2 below)).

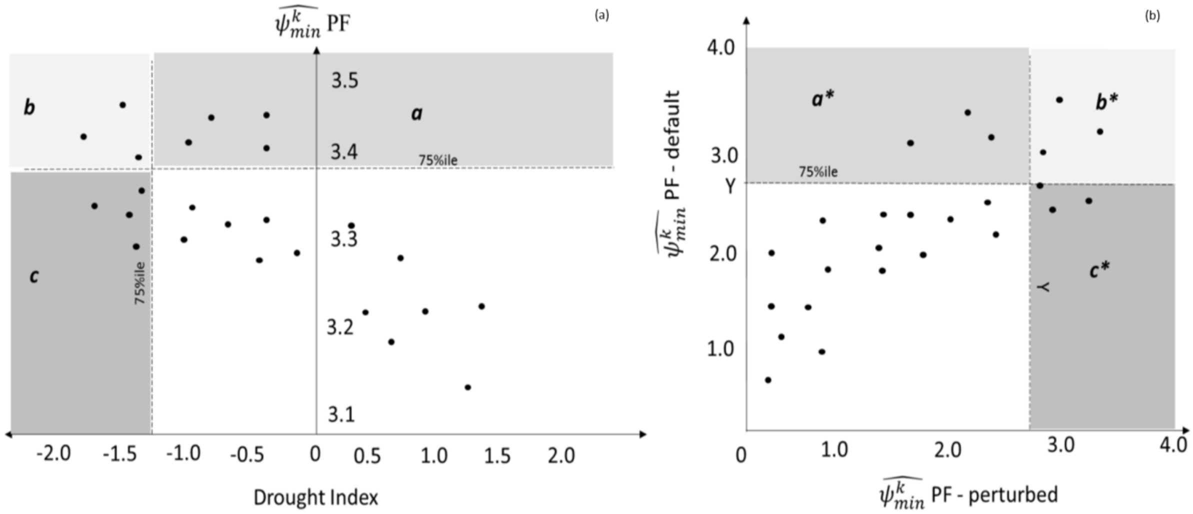

Figure 2: (a) Conceptual schematic of periods of the simulated soil water pressures in relation to the calculated drought index. The threshold values (dashed lines) divide the schematic into four segments, upon which the failure rate (FR) and false alarm rate (FAR) are based on (Section 2, step 3). The segments represent the simulated low soil water pressure events that are not (a) or are (b) detected by the drought index, and the drought events detected by the drought index that do not (c) or do (not labelled) correspond to periods of low simulated soil water pressure. (b): Conceptual schematic of periods of the perturbed Hydrus parameters in relation to the default Hydrus parameters. The threshold values (dashed lines) for default is 75th percentile and the threshold value for perturbed is the soil water pressure value of 75th percentile of default (y).

Note, this is a schematic and the data do not represent the results of this study

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, R^2 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 even a ±10% changes in the van Genuchten parameters disproportionately affect the calculated values (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 large number of graphs (3 sites, 2 depths, 3 model parameters = 18 graphs). Therefore as an example we propose to present one graph in Appendix D – see Figure 3 below.

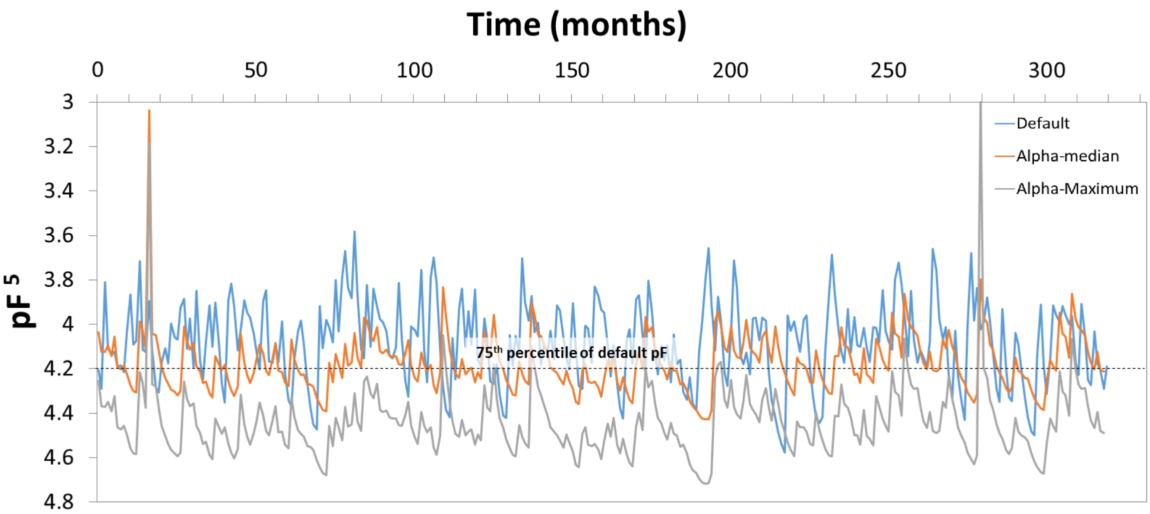

Figure 3. Default and perturbed (median and maximums for parameter alpha) monthly minimum soil water pressure in 5 cm depth for Bourke.

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 not correct. The threshold of the perturbed pF is same as the 75th percentile of the default pF, as shown in Figure 2 and stated in the caption. It would not be a useful performance analysis otherwise. We will emphasise this in the revised manuscript.

Minor comments

1. P1, L6. Replace evapotranspiration with potential evapotranspiration:

   We will replace evapotranspiration with potential evapotranspiration

2. P1, L6. Rephrase as "used as proxy of severity and duration: : :." :

   We will rephrase as suggested.

3. P1, L15. ": : :the frequency with which the simulated: : : below threshold". Actually, you do not want to estimate the frequency, since the frequency is already known as soon as the threshold is defined. Please rephrase.

   Rephrased to: "… and their capability to detect soil moisture droughts potentially critical for plant water stress"

4. P2, L7-8. "…water is controlling…. (e.g., water cycle)". Please rephrase.

   Rephrased to: "… as it is the key variable controlling many processes in biogeochemical cycles such as …"

5. P3, L6. I would rephrase as something like "The analysis …" Since you are not actually strictly describing a "method".

   We will rephrase method to analysis

6. P3, L 27. How many years were used for the fitting? Which period (the full period?). Please clarify. Also, you should say something about the quality of the fittings (the same is true for RDI).

   For both SPI and RDI we used the total length of rainfall/evaporation (table 1) and the quality of the fitting was always R^2>0.97 and RSME = 0.016-0.052

7. P4, L21. Appendix A is just a table. Do you really need an appendix for a table? Same for the other appendices.

In order to maintain the flow and the consistency of the manuscript, we prefer keeping the Hydrus-1D configuration in the appendix.

8. P4, L22. Remove the parenthesis before "Australian" and move it before "2011". Thanks for pointing this out. We will make the change.

9. P5, L1. Please clarify if minimum means minimum among the 30sh daily values in a specific month. Also, please include a standardization of the variable for the successive comparisons with SPI, RDI.
   It's the monthly minimum for both 5 cm and 30 cm soil depths. Detailed explanation of standardization of pF values are provided under responses of reviewer #3

10. P5, L8-16. Please add at least a skill score.
    Addressed under major comment 3.2

11. P5, L17-24. This part on the definition of the threshold is unclear. Please clarify if the threshold is calibrated or not, since you contradict yourself successively in the text. Also, is the threshold computed for each month separately (e.g., 12 thresholds) or for the whole year? The first would be definitely better for the pF.
    Addressed under the responses of major comments 2.

12. P5, L20. In Arnold et al. (2014) is reported that there is still seeding also at the wilting point, which does not means that there is no stress. The capability to germinate is clearly reduced compared to optimal water conditions. You should check Cammalleri et al. (2016) "A novel soil moisture based drought severity index (DSI) combining water deficit magnitude and frequency" where a combination of water stress and frequency is used to define drought from simulated soil moisture. Your definition based only on frequency can lead to erroneous estimates over wet areas.
    We will discuss the reference in the light of our approach and findings. Interestingly, Cammalleri et al. use a complex land-surface model to infer long-term soil moisture data, demonstrating the limited availability of long-term empirical soil moisture data to test the capability of drought indices to detect soil moisture droughts. This encourages us that our approach of using empirically derived soil water retention

curves and a physically based soil water model as a reference or control scenarios is a valuable alternative (Response 1.1).

13. P5, L21. “all values below zero: : :”. It seems that this is not what was done since the 75% threshold is not at zero.

    We agree with the reviewer and will remove that sentence.

14. P5, L22-24. This statement is true only in theory (see Fig. B) and it also highlights how it does not make sense to test two indices if is known a-priori that they would be the same.

    Addressed under major comments 2 and 3.1

15. P6, L9. How do you define the “most extreme droughts”? How was the interval -0.5, 0.5 chosen? Also, please report that 10% steps were adopted.

    Droughts are ≤ -2 considered as extreme droughts (McKee et al., 1993) we will include the reference to the text.L10. We selected ±0.5 assuming the input parameters may vary between this ranges. We will include in the method that we used ±10% steps.

16. P6, L9-14. To compare extreme pF values in different sites does not make much sense, since one site can be “naturally” drier than another which is not related to the occurrence of a drought event (e.g., pF in a dry area after a rainy period can be higher that pF during a drought in Sweden). This is the reason why standardized SPI is used.

    Our underlying assumption is that native plants have been established over long periods and are adapted to the local environmental condition and would suffer similar water stress at the 75th percentile soil water pressure across the three locations. Of course this implies different absolute quantities of soil water pressure. For example, the 75th percentile corresponds to pF 3.4 in Bourke, but is only pF 2.3 and 2.1 in Melbourne and Cairns, respectively (Fig. 5 in the manuscript).

17. P6, L28-29. This contradict what stated in the methodology.

    P 6, L28-29 is under methodology

18. P8, L28-30. The FR alone cannot fully explain the performance of SPI. For instance, is this better that randomly guessing drought events? E.g, How skillful is this index?

    Addressed under major comments 3.2

19. P9, L20. Is more representative compared to what? (I assume to PDSI considering the reference). I would rephrase as "it represent well …".

    It has to be changed as "SPI is representative of soil moisture variation"

20. P10, L9. "rather well: : : then: : :". Please rephrase.

    " … than for deeper soils."

21. P10, 18-19. This sentence is not clear; also, Fig. 5 seems not relevant to this discussion. : A particularly interesting result was that the SPI index, which excludes the effect of PET, performed considerable better than the RDI index (Fig. 5), which may have been due at least partly to the accuracy with which our application of Hydrus simulated evaporation."

    The web plot in the centre of Fig. 5 illustrates that the correlation between RDI and pF greater than SPI vs pF in Bourke.

22. P11, L11-12. This is not necessarily the case. If a significant trend in soil moisture is observed on the site, the use of a longer time-series could negatively affect the analysis (without proper de-trending, etc.).

    We will consider this as a limitation in the revised manuscript. Thank you.

23. Fig. 4. Please report the starting/ending dates, as well as a time scale that is multiple of 1 year (e.g., 12-24, 48: : :) to make the figure more readable.

    We will revise the figure 4 aligned with the editors suggestions and will include further details as necessary.

24. Fig. 5. "The plots represent the highest correlation". What does it means? Please clarify.

    At each location we have assessed 4 perturbations of 2 soil depths and 2 indices. The scatter plots represent the highest correlations between these two variables.

25. Fig. 6. Please re-arrange this figure to make it clearer. E.g., order for site and depth, etc.. Also, the acronym C30, M5, etc. are not defined. Please clarify.

Bars in the figure 6 arranged from highest failure rate to the lowest "The difference between the FR*-FR for all sites and soil profiles for SPI". The C, M, B, are referred to Cairns, Melbourne, Bourke and 5 and 30 refers to the soil depth. We will define them in the figure caption.

References

McKee, T.B., Doesken N. J, Kleist John, 1993. The relationship of drought frequency and duration to time scales, Proceedings of the 8th Conference on Applied Climatology. American Meteorological Society Boston, MA: Anaheim, California, pp. 179-183.

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