Dear Referee,

Thank you for taking the time to read the manuscript, as well as for commenting on parts of it. Below you will find the answers to the comments you made about the manuscript titled "Evaluation of drought representation and propagation in Regional Climate Model simulations over Spain" by A. Barella-Ortiz and P. Quintana-Seguí.

General comments

1. The language throughout makes it difficult to understand at times (for example, P23L18-19: It is not clear if this means that RCMs are not appropriate to calculate SPI at longer accumulation periods) and should be improved dramatically before publication.
   The manuscript will be sent to a Scientific Editing Service to assure that English is correct. Our purpose in P23L18-19, is to point out that the time accumulation period used to compute standardized indices influences the indices' added value. Therefore, we do not state that RCMs should not be used to compute SPI. However, the text will be rewritten to clarify this idea.

2. The paper lacks clearly defined aims and applications, currently it reads as a modelling exercise rather than science to support real-world applications – this can be addressed by a better structured introduction as it currently jumps around without a coherent story (e.g. what is the problem, what have others done in the past, what is the research gap, what is the aim of this study and how this will address the research gap).The introduction and literature review also relies heavily on the IPCC reference, without reviewing peer reviewed publications (and where papers are introduced, they are often listed as 'other papers on the topic' such as P2 L32) and outlining the research gap this paper is aiming to fill. The lack of aims and disjointed nature of the paper make it difficult to reach the conclusions set out in the final section of the paper.
   Drought is a climatic risk, which will become more frequent, severe, and lasting due to a warmer climate. Therefore, it is important to know the evolution of drought. For this, the current modelling tools must be first evaluated. There are several studies about drought representation by models using different types of drought indexes. Our aim is to contribute to these studies by analysing how regional climate models represent drought, as well as the propagation from a precipitation anomaly to a soil moisture and streamflow anomaly. In addition, our study complements a previous one by Quintana-Seguí et al. (2019) which analyses drought representation and propagation by land-surface models using the same methodology as we have employed. As a result, this study improves our current knowledge on the regional climate models' capability to reproduce both drought and its evolution.
   The Introduction section will be rewritten and clarified. We will provide and explain more references to avoid lists. In addition, to increase the focus of the paper on the analysis of the RCMs, we propose to delete results from LSM simulations in the hydrological

drought analysis (Tables 4 to 6: columns “ERA-ORC”, “ERA-ISB”, and “SFR-ISB”), together with the text referring to these.

3. It has not been made clear why this modelling approach was used. The assertion on P6L20 that the atmospheric feedback not being accounted for makes LSMs a good tool to study drought because they can be treated like physically distributed hydrological models, does not provide explanation – why do you want them to behave like a physically distributed model?
   The reasons for using LSM offline simulations as reference in the soil moisture drought analysis, are the following:
   1. Using offline simulations avoids biases due to the atmospheric model and the coupling between LSM and atmospheric model.
   2. There is no observed truth available for soil moisture (P7,L1-2) that covers mainland Spain, thus we use offline simulations as a reference.

   We will rewrite the text in P6L20 to explain clearer that atmospheric model biases are excluded in LSM offline simulations.

4. You don’t mention or address the issue of uncertainty – what about the uncertainties of the modelling approach? Could you explore this using a multi-model ensemble?
   Yes, uncertainty is a very important topic and, in fact, we are already dealing with it. The main objective of this article is to evaluate drought properties in RCM simulations. We are performing the analysis on three different models and these provide different results in terms of drought indices and, specially, drought propagation. With only three models we already show that the values of n_x (time scale of propagation of drought) are very different for both soil moisture and runoff. Adding more models would show the spread with more detail, but we believe that the differences with three models are large enough to show that RCM developers should look at these issues and RCM users should take them into account.

5. In many places, there is text seemingly in the wrong section of the paper, for example P3L33-P4L4 should more likely sit in the data/methods section as this is detailed for the introduction. P5L1-8 would be better placed in the introduction. P11L24-27 would be better placed in the discussion describing why there were discrepancies between the modelled outputs. The authors should review the text to ensure that descriptions of data and methods and discussion text are in the appropriate sections.
   - P3L33-P4L4 should more likely sit in the data/methods section as this is detailed for the introduction. In our opinion it is not too detailed for the introduction as we list the RCMs analyzed and the references used. These are further explained in Section 3. However, if the referees and editor deem it necessary, this part of the text from the Introduction section can be shortened.
   - P5L1-8 would be better placed in the introduction. The text describes drought in mainland Spain, which is the area of study. That is the reason why it is located in

Section 2 (Area and study period). However, it can be moved to the Introduction Section if the referees and editor deem it necessary.

- P11L24-27 would be better placed in the discussion describing why there were discrepancies between the modelled outputs. We propose to move the phrases " There are few pluviometers in mountainous areas, probably causing an underestimation of precipitation over these regions. However, this effect is limited to mountainous areas" to the Discussion Section. However, we believe that lines 26 and 27 should remain in Section 5.1.1, because they describe CCLM4 mean precipitation over mainland Spain.

6. Section 3.1: this should have more introductory information before diving into the detailed descriptions of SAFRAN and ERA, in 3.1 please outline what variables you use and what they are needed for before describing them in turn.
   More introductory information will be provided before the subsections describing SAFRAN and ERA-Interim. Besides referring to Table 1, we will explain why these were chosen as driving data and reference dataset.
   The variables used in our study to compute standardized indices (precipitation, soil moisture, streamflow, and total runoff), are detailed in Section 4.1 (Drought indices calculation) (P9L14-16), not in Section 3.1. However, we do provide a description of SAFRAN's precipitation in Section 3.1.1 because it is considered our reference dataset for the meteorological drought analysis.

7. In Section 3 the RCMs are introduced third but surely start the modelling chain, I suggest you introduce these first, then the LSMs then Hydrological Models. Was it necessary to calibrate and validate your models – how did you do this?
   Yes, we can clarify this point. You mention "modelling chain" as if the RCMs were driving the LSMs (one way forcing). We fear there is a confusion here. The RCMs contain themselves a LSM, which is coupled with the atmospheric model of the RCM. We are taking the outputs of the RCM's LSM variables directly from the MedCordex database. Thus, there is no "modeling chain" in this regard. This is, we are not forcing standalone LSMs with the outputs of the atmospheric variables of an RCM simulation. We also use standalone offline LSM simulations, forced by ERA-Interim and SAFRAN, in order to have comparison points.
   That being said, we propose to restructure Sect. 3 as follows, in order to provide more clarity:

   3 Datasets
   3.1 Forcings and driving data
   3.1.1 SAFRAN meteorological analysis
   3.1.2 ERA-Interim
   3.2 Models
   3.2.1 Regional climate models
   3.2.2 Land Surface Models
   SURFEX

ORCHIDEE

3.2.3 Hydrological models

3.3 Observations

SURFEX and ORCHIDEE sub-sections will not be numbered, because only three levels of sectioning are allowed.

Concerning model calibration, the situation is as follows:

1. We took RCM data from the MedCordex database. RCM modelers do tune their models, but this information is not available to us. In this regard, we simply are users of the RCM outputs.
2. We used an ERA-ORCHIDEE simulation, which was provided to use by Jan Polcher (IPSL). We do not know how IPSL calibrates its LSM.
3. We performed ERA-SURFEX and SAFRAN-SURFEX simulations. We did not calibrate SURFEX. We used default values for all non physical variables (i.e. subgrid runoff). The corresponding flows were calculated using the RAPID river routing scheme. The Muskingum parameters were not tuned, we used default values too. Concerning SURFEX modelled flows, we have proposed in comment number 2 to delete results from LSM simulations in the hydrological drought analysis.
4. We also used the outputs of the SIMPA model, as a reference. This model is heavily calibrated. The model calibration and run were performed by CEDEX, the Spanish institute that provides the reference natural streamflow simulations to the Ministry for the Ecological Transition and the basin authorities. We are users of these simulations and do not have access to information on the calibration.

   This information can be included in the manuscript if the referees and editor deem it necessary.

8. A lot of detail is provided about the LSMs which is published elsewhere and appears to pad the paper, much of the model background can be removed – the focus should be on why the models were chosen and what they will be used for.

   LSMs' descriptions will be shortened and the reason why they were chosen and the use they were given in the study will be explained.

9. Section 3.4: It would be useful to include a map of the observation stations used- how many stations were used? Only 8 across the whole of Spain? Why not more – there must be more than 8 stations that have 95% data completeness?

   You are right, there are more than 8 stations with 95% data completeness. However, our criteria to select them was more demanding:

   - 95% data completeness: This assures that the observation monthly series have few gaps.
   - Area greater than 10^3 km^2: Since streamflow is approximated by runoff, it is likely to perform poorly in small basins considering the coarse resolution of the RCM simulations. Therefore, the analysis was limited to large areas.

- KGE between SIMPA and the observations greater than 0.5: To consider a near-natural regime.
  These criteria are explained in Section 5.3 (P18L1-11). The second one was the most restrictive, since only 13 stations out of the 87 with 95% data completeness have an area greater than 10^3 km^2.

10. P18L8: What evidence or scientific literature did you use to select the 'arbitrary' KGE of 0.5? Later in Section 5.3.1 you say performance of CL4 is poor because KGE is generally below 0.5, but the best performance is for RS4 with KGE of 0.7 – is this enough of a difference between poor and best (reading 'good') performance?
    - P18L8: What evidence or scientific literature did you use to select the 'arbitrary' KGE of 0.5?
      In order to validate the aggregated runoff of the RCM simulations, we needed gauging stations that were as natural as possible and whose corresponding basins were large enough to be compared to a low resolution RCM. This is difficult in Spain, due to the high degree of human influence. Thus, we needed to have enough large basins that were as natural possible. We thought that a high value of KGE between SIMPA (naturalized flow) and the observations was an indicator of natural regime. Then, we had to set a threshold of the KGE. We tested different values and 0.5, was a reasonable compromise between "near natural regime" and "enough number of stations". It is true that 0.5 is not very high, and thus some human influence can be present, but we had to draw the line somewhere and we did a sensitivity analysis based on our own judgement. Thus, the value is not as "arbitrary" as the text implies. We propose to remove the word "arbitrary" from the text, as it is misleading, and clarify the procedure we followed to select this value.
    - Later in Section 5.3.1 you say performance of CL4 is poor because KGE is generally below 0.5, but the best performance is for RS4 with KGE of 0.7 – is this enough of a difference between poor and best (reading 'good') performance?
      Regarding the difference between poor and best, the text in Section 5.3.1 does not refer to a "best performance":
        - It compares RCMs with LSMs and states that two RCMs provide better KGE than the LSMs with the same surface scheme (P18L19-20).
        - We say that CCLM4 behaves poorly because we have set a minimum threshold of 0.5 and all KGE values provided by the comparison of CCLM4 and SIMPA are lower.
        - When RCSM4 and PROMES are compared between them, it is said that RCSM4 shows the best KGE value for one station, but that PROMES behaves better over both basins (P18L22).

      We propose to modify the text so that it is clearer that CCLM4 provides the worst performance of the three RCMs analyzed and to replace " PMS behaves better" by "PMS performs better". In addition, we will explain that PMS is identified as the best performing RCM out of the three RCMs analysed according to the KGE

average value. In this section, We will also avoid using adjectives such as “poor” to qualify the KGE values, sticking more to the numerical values, in order to avoid misleading the reader with our own subjective views.

11. Section 5.3.1 – why were the temporal analysis not shown? If you only have 8 gauging stations, it would be simple to include time series plots showing the modelled ensemble data against the observations.
    Temporal analysis are not shown in order to reduce the number of figures. The main result of this section is Table 4, which compares streamflow and aggregated runoff to analyse the performance of RCMs by means of the KGE. The temporal series provides additional information, which in our opinion is not necessary to include. However, if the referees and the editor deem it necessary, we can include them.

12. Table 6: This might be better as a figure with the catchment areas coloured by SPI-nx – as readers we don’t know where your catchments are, how do the results vary spatially? What might the effect of catchment properties be on the propagation process? How well do the different models represent these catchment properties?
    We thought about making a similar figure, but we discarded it. The figure we planned to make was to plot the points of the stations, graduating their color in function of the value of n_x. However, as we are using only 8 stations (because we want to compare the simulations to observations from near natural flow gauging stations), the maps were not really necessary (a lot of empty space, with just a few colored points). You propose to color the areas of the basins. This can be done, but, again, with only 8 stations, the resulting colour areas would not be meaningful enough. However, we propose to include a figure showing the relief, river network, and the ubication of the gauging stations. We will include, if possible, the catchment areas defined by each station.

13. Table 6: What r values are associated with the Evans classification? How significant are these correlations? The bold type face mentioned in the caption is not obvious in the table.
    - The correlation ranges from the Evans (1996) classification will be included in the caption.
    - Correlations are 95% significant. This information will be included in the manuscript.
    - Scales longer than 12 months will be marked in bold.

14. P22L4: why do you believe standardised indicators are appropriate for this study? This should have been outlined previously.
    Standardized indices define drought according to the variability of a given variable (P4L6 and P9L9-10). They allow to study different types of drought depending on the variable selected. In addition, variability is very important regarding drought analysis and it is the basis of these indices. We propose to extend the text to make clearer why we used standardized indices.

15. P22L10: what is meant by event extension? The duration and intensity (and extension) of events is not described elsewhere nor shown in any figures – what do you refer to here?
    By “event extension” we refer to the area affected by a drought event. We will rewrite the text to make it clearer.
    The duration and intensity of events is shown in Fig. 2, while the extension (area) is treated in Fig. 3.

16. In general the figures were too small and labels too small to read. You should avoid the red-blue colour schemes of Tables 3-6, they are not appropriate for those who are colour blind. You can check whether your figures are colourblind friendly here: http://www.color-blindness.com/coblis-colorblindness-simulator/ or by using the CVSimulator app
    We will increase the figures’ size. We will also check the tables in the Color BLIndness Simulator and change the colour schemes to make them more appropriate to colour blind people. Thank you for pointing this out.

17. In regards to the Barker et al. (2015), you have cited the Discussions paper, please cite the final 2016 paper (https://www.hydrol-earth-syst-sci.net/20/2483/2016/hess-20-2483-2016.html).
    Thank you for pointing this out. We will cite Barker et al. (2016).

18. On P3L31 Lopez-Morreno’s name has been misspelled.
    Thank you for pointing this out. We will correct the reference.

19. The tense throughout is the present tense, however, research is conventionally written up in the past tense (as it is work that has been completed), please correct this in the next version of the paper.
    We propose to write the state-of-the-art from the Introduction section, as well as the Methodology section in the past tense and leave the results description in the present tense.

We hope you will see a clear improvement in a revised version of the manuscript.

Yours sincerely,

Anaïs Barella-Ortiz