Review of :
egusphere-2023-119
Extratropical circulation associated with Mediterranean droughts during the Last Millennium in CMIP5 simulations
W. M. Kim, S. J. González-Rojí and C. Raible

Evaluation:

The purpose of the manuscript valid and very interesting, and the methodological approach is reasonable for the aims of the study. The data used also represent a variety of sources that can bring confidence to the assessment. I think the study will eventually deserve publication. In its present state I would like the authors to consider a number of arguments that are listed below. Quite a few are methodological and needed for a proper understanding of the results. Also, they are needed to properly assess the correctness of the results. It is possible that the results will not change much after considering some of the suggested methodological issues but I think that at least the manuscript can improve from making them more clear.

GC1. Abstract:

I think the abstract is in general well written and it emphasizes the findings of the manuscript.

GC2. Introduction:

I think the introduction does its job in representing the frame of the problem and the case for this study.

There are a couple of issues I would like the authors to consider though:

1. Page 1, lines 21-23. 'The climate of the…'
This paragraph divides the seasons into wet and dry and refers to the wet one specifically as the winter season. There is considerable spatial variability in the rainy seasons in the Mediterranean lands and for many, the wet seasons are spring and autumn. I think this winter/summer separation may be misleading. Consider for instance Xoplaki et al (2004; DOI 10.1007/s00382-004-0422-0) or others, where an extended wet season is used. I think this discussion should incorporate better this seasonality character.
Note page 2, line 24 where you mention 'highly variable seasonal hydroclimate...', although the end of this paragraph (line 30) will make the emphasis again on winter.

2. Page 3, lines 64-65. 'They also found that … is model-dependent'.
This seems to relate specifically to the results of the current manuscript and perhaps should be discussed specifically either here, providing more specific background and describing in which way this can happen, or/and

at the end of the manuscript discussing what this manuscript provides in the context of previous evidence.

GC3. Section 2: Data
The data section can perhaps be improved by discussing a bit more the model sources and the rationale for decisions concerning soil moisture data, as indicated in the following.

1. Section 2.1 Page 3, lines 83-86.
   The CMIP and PMIP and other model efforts could include a reference. I am aware they were just provided in the previous paragraph, but it seems pertinent to me that they get those also in the Data section. I leave this to the taste of the authors.

2. Section 2.1 Page 3, lines 85+ ‘We consider only simulations…’
   There are some issues that can be considered related to the use of soil moisture in different models:

- The models have very different depths and therefore they will possibly produce different soil moisture statistics, even if depths only down to 0.7 m are considered. Using this set up of models is fair and possibly it contributes to inter-model differences, but also it is arguable that models with shallower depths may be less realistic. Note that some models like GISS only have a depth of 3m. This will produce potentially a different vertical distribution of moisture and different temporal variability. Perhaps it is something the authors would like to comment on.

- In Table 1 it is included that a depth of 0.7 m is considered. I think this is relevant and should also be included in the text. It would be desirable to include a rationale for this decision and how it may influence the timescales of variability of soil moisture.
  What is the typical range of depths for the bedrock, that limits the presence of water, for the area of interest in general in these models? Including a shallow limit may be more representative for the whole area but also exclude lower frequency variability more typical of deeper levels. I think including a rationale for this in the text of Section 2.1 would be good.

3. Section 2.2
   Why are NOAH-LSM data used for soil moisture and ERA5 only for circulation? I mean what does NOAH-LSM offer that you would not get in this manuscript from ERA5 soil moisture?. There is probably good reasons for it. I would just suggest that a motivation for the use of these data, beyond the fact that soil moisture observations are scarce, is provided.

   Also:

- Line 87 ‘… analysis: Giss-E2-R, CCSM4…’
  References should be provided for these simulations. Considering previous points, I would suggest that a minimal background is at least provided in how different the soil moisture modelling can be among all

models. This would help giving the reader an idea of what to expect in terms of inter-model differences. For instance, vertical resolution can be very influential for soil moisture, as indicated in line 90…

- Line 87 '… 12 ensemble members of CESM1'
  There are 13 available if I am not wrong. Nothing critical but maybe you want to state why that selection.
  Here it can also be stated this refers to the all forcing simulations, although it is quite clear in the context (see next).

- Line 91 ' All simulations were run with the volcanic, solar…'
  It is clearly stated earlier in the introduction that this is all about internal variability. Perhaps some comment about focussing on forced simulations and not in long control runs (also available in cmip) is pertinent.

- Line 101 ' …has four layers up (down) to two meters'
  Ok, but I understand that only the soil moisture of the first 0.7 m is used, right? Explaining this in relation to the previous information in Section 2.1 would be clarifying.

GC4. 3. Methods
3.1 Calculation of anomalies.

I find the definition of annual anomalies and the explanations for it somewhat confusing. There is an emphasis in this explanation that I do not really understand. Perhaps this is my fault but, please, consider it and see if this influences the text as it stands now.

1. Section 3.1 Page 5, lines 111+.
   - 'These variables are transformed into the annual mean anomalies at each grid point'
     At this point I would say it is not possible to understand what is meant if additional information is not provided first, and I would argue to simply write '…transformed into annual anomalies'. I think that will be enough also for the rest of it.
   - 'We use annual mean anomalies in order to include winter conditions in the analysis, as it is an important season for the annual hydroclimate in the Mediterranean'
     I think it is important to discuss this in the context of what is indicated in GC2.1
     Specifically for this sentence, it is not possible at this stage to understand what is meant. Not a big thing because one can perhaps understand it from the following, but:
     a) I would suggest to include the description for the intra-annual anomalies before and then explain they are averaged; if these arguments stay in the text.
     b) I would actually (respectfully) challenge that this is true. I think it is not, but maybe I understood things wrong. Se below.

- ‘Prior to the anomaly calculation … to the 70 cm level’
  These two sentences are relevant and the reader could go with knowing this earlier in the data section (GC3.2).
- ‘Then, first, the monthly anomalies… monthly values’ […] ‘Second, the necessary annual mean … these monthly anomalies’
  I do not understand these two sentences. If you calculate monthly anomalies (differences between monthly data and the long-term monthly mean you are essentially subtracting the annual cycle. If you average out the resulting monthly anomalies what you get is the annual anomaly, equal to calculating anomalies from annual data by subtracting the long term mean.
  Therefore, I do not understand the previous sentence about the importance of winter or the unnecessary emphasis on the monthly anomalies if you end up in annual anomalies.
  The sentences in the middle about reference periods read fine to me. Perhaps it fits better with the next paragraph and the issue of the trends since their definition is somewhat related.
- ‘Thereby, the ensemble means of the anomalies… Maher et al (2018)’
  With the preceding sentences I am confused. I understand the purpose is subtracting the ensemble average to each member to get rid of the long-term trends; or better said of the forced response in general as it will also affect volcanic events for instance. This would be consistent with the following arguments in the paragraph. Consider explaining/rephrasing these paragraphs.
- ‘This method guarantees that only internally driven variability remains in the time series of the variables’.
  … mostly internally driven variability, but only in the post 1850 period. This should be perhaps indicated specifically.
  However, what I miss a bit is the rationale of why it is done like this. Why is the forced signal intended to be filtered out in the post1850 (natural and anthropogenic) and not in the pre-1850 period? The reader has to guess that perhaps this is intended to get rid of the long-term trends, and the rest of the forced signal goes with it

- Note also that the ensemble is calculated over different models in CMIP5-PMIP3 if I understood well (?). Now this is an issue that can be relevant, as different models will show different levels of response to external forcing. When this is done over an ensemble with the same model you can reduce the external response in each model run as it is assumed that the external response is common to all runs. One aspect here is that depending on the number of runs you are also subtracting also internal variability that has not been averaged out in the ensemble average, and this is a limitation that can be discussed. A more relevant aspect can be that if you are using different models in the ensemble as it is the case, the assumption that a common response to external forcing is filtered out does not hold and by subtracting from each model the ensemble average you do not actually filter out the external response of that specific model.

Therefore, this would work for the CESM-LME ensemble but not if you mix in the ensemble different models. This can be a sensitive issue that should be at least clearly discussed here or indicated here and better discussed later in the text.

- The two periods, LM and Hist, … models and regions'

  The pre-1850 and post-1850 are merged. However, they do not have the same nature, as in the pre-1850 the natural externally forced variability exists and in the post-1850 it is not intended to exist. Although if calculations have been done for some models out of a multi-model ensemble, some leftovers of externally forced variance should remain. Nevertheless, this should be in principle not important if externally forced signals are not pursued and are not expected to play a role either; actually, for that purpose, control runs could be as adequate or more adequate.

  I am trying to highlight with the previous sentences that some level of confusion can be transmitted to the readers.

I think the last paragraph is a good example that this section could benefit from some rethinking. I would say that not just in the details of the anomalies, but it is also important to provide a rationale for the reader to understand the strategy of the study… its advantages, and its limitations.

2. Section 3.3 Page 6.

   I wonder why not considering the soil moisture from ERA5 or ERA5-land additionally or instead to NOAH-LSM. This relates to GC3.3. Since ERA5 produces soil moisture out of its assimilation system, this would be physically consistent with other ERA5 variables used here. Perhaps the authors want to make some comments in the text about the advantage of using the NOAH data here instead.

3. Section 3.4 Page 6, lines 161+

- 'When these temporal and spatial …'

  The regional arguments about drought incorporate the condition of having 60% of the horizontal gridpoints with negative SOIL. This faces some difficulties with the fact of using different horizontal resolutions. However, this could be overcome by considering the spatial size of anomalies by using grid box area and the total spatial extension that the anomalies represent, instead of the number of gridpoints.

- 'Next, the weighted spatial average of SOIL is … Mediterranean'

  How is this weighting done? Perhaps I missed it earlier in the text?

  If this is done, why not using this regional average to determine the occurrence of spatially large enough droughts?. I do not think the result will be very different from the previous approach and it would overcome the use of a percentage of gridpoints with different model resolutions.

4. Section 3.4 Page 7, lines 171+
   ‘Only climate models with some ensemble members … not detectable’ This is already a report of results. I have no major concern with anticipating it, but what is the reason for it?. The sentences are descriptive of the fact that pan Mediterranean droughts happen in some simulations with an specific feature and not in others. Can you at this stage argue about this feature?, or should this moved further down in the text and a discussion provided?

5. Section 3.5 Pages 7-9
   I get the structure of the methodology in general, however I suggest the authors revise this section for a more clear explanation, revising notation and perhaps the current state of the explanation of details of the methods that can be more important for the understanding of this section. I will not go to details in an exhaustive way, just provide some examples.
   The text is providing the sequence of a method. How the outcome of each of the steps feeds the following steps should be clear in explanation and, I would suggest, a homogeneous notation. I also recommend there is a rationale/justification for the conceptual use of each step. I will briefly try to highlight this with examples on the following, but please, go beyond those.

   References: the authors provide reference but please take care they are appropriate. For instance, there are many ways of applying PCA and the reference to the correct texts that describe the approach presented herein should be provided (see below).

   Use of maps and series: This should be clear from the text. For instance, in the PCA, I understand T(t) are the principal components. Please, indicate that and also the range of the parameter t. The same with s, it is good to indicate the range so that the number of modes or the retained number of modes is well defined

   In equation (1), where are the eigenvalues?. I understand they are multiplying either the spatial or the temporal component. They should be indicated. I understand it is particularly relevant if they multiply the temporal component. The reason for it is that the temporal component will have a standard deviation 1 or different from it and this will impact the Kmeans procedure as it will affect the distance.

   Explained variances are mentioned but it is not said that the 70% reflect accumulated variance accounted for by a number of pcs.

   Some sentences are confusing. E.g. line 191 ‘PCA is applied to the Z500 fields during droughts for each model’. I do not understand this.

   How do the PCs enter the following analysis? They have unit standard deviation (perhaps not), and their inter-pc correlation is 0. How does this play a role in the Kmeans clustering, what does it mean physically because pcs should indicate different modes in time that are afterwards grouped, even if they are uncorrelated.

The notation of how the T(t) go into equation (2) should be consistent with this equation. I suggest that the notation is blended for the various steps of the analysis. If it is not done, readers will have to accommodate how things fit from their knowledge and from the different steps of the method. If notation is blended, this section would actually describe one thing, the approach followed in the manuscript, not several independent methods. Some features of notation are repeated for different things, e.g., s as parameter for points in space and the Silhouette coefficient for each point.

There are 71 clusters. I learned this from the text but realized that this is the sum of all clusters from different models in Table 2. For each model 5 to 6 pcs are retained and from this, between (mostly) 3 and 6 clusters are formed. What does this mean? Are the 3 clusters gathering the information of the 5-6 pcs? In which way? Some rationale/explanation for what is conceptually happening is good for the reader.
Line 220: at this level I do not know what the correlation between clusters means.

All in all, section 3 needs, in my opinion, to be well revised do deliver a more clear and consistent text. This does not necessarily imply changes in the calculations, nor the results of the following sections. But it may impact (positively) the interpretation.

GC5. 4. Results
4.1 Observation-model comparison

1. Related to previous comments:

   Line 237: '… by subtracting from each of the ensemble members the anomalies at each grid point'
   This relates to previous comments and could be explained better.

   Line238: 'The spatial correlations between SOIL and Z500 of NOAH-LSM and each of the climate simulations are presented in Fig. 3'
   I do not think it is wrong at all but I wonder what is the gain of using NOAH instead of the soil moisture from ERA5 in this figure.

2. Figure 2 and related
   2b and 2c are good in showing the impact of using different references. However I would say that the two are not really needed. With one of them it would be enough to explain it. Perhaps the space could be saved to accommodate a time series of the full 850-2005 period. I leave it for the authors to decide.

The shading in Fig 2b seems to indicate consistency in the range of variability with the range of 'observed' NOAH-LSM variability, which is good to indicate.

3. Figure 3 and related

- Watch some statements like ' … negative correlations over southern Europe, but the correlations outside Europe are not significant'
  They are for CCSM4 and bcc in the western low latitudes of the domain.

  Line 255, '… all models present similarities to the NOAH-LSM, fed…' What we see in Fig. 3 is the result of both NOAH-LSM and ERA5, right?. Actually, it is likely that the large scale structure we see there is more dependent on the global model; one could actually test if it changes with other reanalysis products. I think this is likely out of the scope of the study, but it is not totally off the line of argumentation because this correlation field with ERA5 is what we consider 'truth', but it could change if we would have used a different reanalysis product.

- Perhaps a more relevant issue: if you consider the variability among patterns in Figure 3, what is the variability among simulations of one single model ensemble?. I would assume that it is smaller, but it may be worth reporting.
  In the actual figure 3, for those models with ensembles, is the pattern that is shown the result of one single experiment? If so, I think it should be clearly stated in the caption and in the text. I would advise against including correlations using ensemble averages, but I don't think this is what is being shown.

4.2 Mediterranean drought …

1. Figure 4 and related
   In relationship to the role of external forcing, or the lack of it ('This fact emphasizes that external forcing signals do not play a role in droughts over the Mediterranean …' line 286), there are several issues that may be worth commenting.

- How are droughts calculated over sub-ensembles of simulations (GISS, CESM1) in Fig. 4, and how is the ensemble spread provided for them? This should be explained (sorry if I missed it) in the text and figure caption. Also, I expect it will justify the different temporal resolution of the curves in Fig 4 for GISS and CESM1 in comparison to the others. However, it can be misleading as the results of those two models may be read as if soil moisture for CESM1 would be consistently higher (smaller) in the western (eastern) Mediterranean during the late 17th and 18th century, or in GISS also during the 18th and early 19th centuries… or in GISS opposite to that during the 17th century. This would not be possible and would contradict the first statement of no role in external forcing as it is very unlikely that different model runs of a sub-ensemble with different initial conditions will coincide systematically in simulating relatively dry or

wet periods unless forcing would play a major role. The only reason for that would be external forcing.
Also, the finding of opposite phasing between west and east is interesting, and I would argue that it should be more perceptible for GISS and CESM1 if individual runs are considered, in relation to the statement '…observed, more clearly in those models and periods with one ensemble member' (line 296). Therefore, I suggest the authors revise how the ensemble behavior is presented for those two models.
For instance, the last statement of Section 4.2 is sensitive: 'For those models and periods with more ensemble members… sometimes this association is blurred…', it should indeed be the expected behavior, even more than what is shown. If you resort to individual simulations, this should be more clearly evidenced. The ensemble spread should expectedly blur everything since a dry or wet century in one run should not be expected to be consistently dry or wet in most of the other sub-ensemble runs.

2. Figure 5 and related
   I am not against showing temperature anomalies in association to the Z500 patterns, but why not showing precipitation and actual drought patterns. How do the geopotential anomalies account for drought occurrence?
   Figure 6: I need a better description of the methods section to better interpret results and figure out whether they can be dependent on methodological choices.

Minor comments:

MC1. Introduction, page 1 line 19
' Ocean, teleconnections and large-scale modes of variabity.'
Ocean, and teleconnections with other large-scale modes of variability (?)

MC2. Page 2, line 40
'…change in climate boundary conditions…'
'… change in external forcing…' seems to me better in this context. The current sentence is maybe a bias from climate modelling

MC3. Page 2, line 42
'Several natural proxy-based …'
Natural meaning?

MC4. Page 2, line 44
'…summer dry and wetness…'
Dryness?

MC5. Page 2, line 42
'…of droughts variability…' → of drought variability

MC6. Page 5, line 111
'…at vertical soil layers…'
You mean perhaps, vertically integrated soil moisture content?, or something of the sort…

MC7. Page 7, line 172
'…show few numbers of…'
Cases of?