Dear editor Dr. Goosse and the reviewer,

We would like to thank the editor and the reviewer again for dealing with our manuscript and for his/her constructive comments. We sincerely appreciate the time and effort the reviewer dedicated to reviewing our manuscript. Here, we provide our responses to the reviewer's comments.
Our responses are in blue font, and the line numbers are within [the brackets].

Response to the reviewer:

Evaluation:
The authors have included responses and changes to the comments. I am satisfied with most of them. I am not convinced about a couple of them that I mark with (*). I do suggest improving the description of the methodology. Additional to that there are several points which I think should be considered. I regard them as minor and easy to do, so I leave it to the Editor whether I should see the responses and the corrected manuscript again.

**GC1. Introduction:**

I think the introduction has improved with the changes implemented. A minor comment:

1) 1. Page 2, line 51. '... no substantial change in external forcing...'
This statement can be missleading. I understand in the context of what the authors want to convey but I suggest rephrasing. Forcing changes are substantially smaller than say the glacial-interglacial transition, but still detectable in reconstructions and substantial enough to be climatically relevant. Consider rephrasing. I think the relevant issue here is that the LM is the period in which the (natural) forcings are most comparable to nowadays and the one, if we think preindustrial, that offers the reference before the development of the large anthropogenic forcing.

Thanks very much for the point. We modified the sentence to:
[51-52] "the last millennium is particularly interesting as it is a relatively close period with similar natural forcing conditions to those during the pre-industrial era"

2) 2. As a reference for drought variability at global scales during the last millennium and potential responses to internal variability and external forcing consider https://cp.copernicus.org/articles/16/1285/2020/.

Thanks for the suggestion. We added in the introduction
[71-75] "the impact of volcanic eruptions on hydroclimate-related variables and long-lasting droughts has been assessed on a global scale (Stevenson et al., 2017; Roldán-Gómez et al., 2020) and the Mediterranean region (Kim and Raible, 2021) using the climate simulations from the Community Earth System Model (CESM; Lehner et al., 2015; Otto-Bliesner et al., 2016) and the fifth phase of the Climate Model Intercomparison Project (CMIP5; Taylor et al., 2012) - Paleoclimate Model Intercomparison Project Phase 3 (PMIP3; Schmidt et al., 2012)."

[78-79] "On a global scale, Roldán-Gómez et al., 2020 indicated that there is not a clear relationship between soil moisture and external forcing that can be detected during the last millennium in the CMIP5-PMIP3 simulations.

GC2. Section2:Data

3) 1. Page 4, line 99: (LM; 850-1850). See consistency with pate 5 line 140: 850-1849
We changed it to 850–1849, also in other parts of the text. [103]

4) Page 4, line 109: '...are transient ~~that~~ and were run...'
We changed it as suggested. [113]

GC3. Section3:Methods

5) 1. Line 142: '... the effects of GHG on drought ~~is~~ are not...'
We changed the sentence to:
[144-145] "This means the effects of increased GHG on droughts are not included in the analysis."

6) Line 167: '... present-day NOAH-LSM-ERA5 correlation field and that obtained from the climate models ...' (?)

We modified the sentences for clarification:
[167-171] "In addition, to quantify the spatial similarity of the correlation fields obtained after the PCC between the observation-based data (NOAH-LSM--ERA5) and the climate models, the pattern correlation is calculated between the two datasets for the western and eastern Mediterranean regions. For this calculation, the horizontal resolutions of all data are interpolated to match those of the coarser climate models, which are bcc-csm1-1 and MIROC-ESM (Table 1). The pattern correlations assess the overall resemblance between the correlation field of NOAH-LSM--ERA5 and the climate models for the present day. "

7) Line 175+ Drought definition. It is not clear if the drought is composed only of negative SOIL values (as it is literally stated) or if it contains the dates of the two positive values at the end ('... and continue until two consecutive years of positive anomalies...'
'.. of only negative values without being interrupted by a particularly wet year in between...' However, I understand from the definition that if there is a positive SOIL anomaly, the drought continues until two positive anomalies are found' – I am only trying to point out that I find the explanation confusing in this respect.

Thanks for the point. We modified the paragraph for clarification to
[178-184] "A drought commences after two consecutive years of negative SOIL and continues until two consecutive years of positive anomalies occur (Coats et al., 2013). These two wet years are excluded from the drought period, ensuring that droughts consist only of negative SOIL (Kim and Raible, 2021). This definition guarantees the intensity of droughts by considering only negative SOIL without interruption by a particularly wet year in between. It also assures a minimum drought duration of two years and at least two wet years of separation between drought events."

8) Line 183 '... a substantial portion of the region experiences drought conditions …'
We corrected it. [188]

9) Line 186. 'Nevertheless, this approach avoids changes in the initial SOIL values, as ...' I do not understand what is meant here.
We wanted to mention that we use the original SOIL values without interpolation. We modified the sentences to:
[189-191] "At this step, we do not apply any horizontal interpolation in SOIL. Thus, regional coverage (geographical extension and number of grid cells) differs slightly between the models (as shown in Table 1). The reason is that the hydroclimate variables associated with precipitation can be sensitive to the horizontal grid resolution. "

10) Line 198+ 'Also, the time series of SOIL is generated by applying spatial weights to the soil moisture anomalies, taking into account the spatial extent of each grid cell within the confined region'. This is not clear to me and I think it should be explained, what is done, how is it taken into account. It is pointed out that it is somehow taken into consideration, but I suggest to indicate how.

We simply wanted to say that we generated the spatially weighted time series of soil moisture anomalies for each of the regions. We made the sentence simpler to reduce redundancy to:
[204-205] "For the latter, we generated the spatially weighted time series of SOIL for each of the regions."

Also, we moved it to [204-205] after introducing the wavelength analysis, as the time series of SOIL is principally used for this analysis.

The formula that we used to calculate the spatially weighted time series of SOIL for each of the regions, west and east, is

$$SOIL_{mean} = \sum_{i=1}^{N} SOIL_i \, cos(latitude_i) \; / \; \sum_{i=1}^{N} cos(latitude_i)$$

where N is the number of grid points in the confined regions.

We did not add this formula in the manuscript, as it is rather technical and is available as an embedded function in many programming languages.

(*) Line 214+, PCA calculation
11) Line 222: '...PCs, represented as ...Ti(t) * ui(l)
In my understanding this is not correct. Please, indicate which variables are the PCs and which the eigenvectors, also indicate where the eigenvalues are included in eq (1).
I think this part of the methods is not satisfactory. The explanation should state how the covariance (or correlation?) that is diagonalized is defined (in space or time) and then which ones are PCs, eigenvectors, EOFs, eigenvalues, etc, in a clear way. I think this influences the interpretation of the results and helps reproducibility.

We included the details pointed out by the reviewer, and corrected the paragraph and the eq 1 in the revised manuscript:
[219-225] "Given a spatiotemporal field U(t,l) where l is the spatial dimensions (latitude x longitude), and t is the time steps in years (t is the total drought years), the PCA decomposes the field into M number of modes or principal components (PC) according to the following equation

$$U = \sum_{i=1}^{M} \lambda_i . a_i x_i^{\mathrm{T}}$$

where $a_i$ is the i'th standardized PC of the U data, the $x_i$ is the i'th empirical orthogonal function of the original data (also the eigenvector), and $\lambda_i^2$ is the corresponding eigenvalue that represents the explained variance of the i'th PC ($a_i$). The resulting EOFs are orthogonal, and the PCs are uncorrelated."

12) Why 70%? Likely arbitrary but I would expect the text to say that the results are not very sensitive to this decision.

The 70% value was chosen based on the Silhouette coefficients (S) calculated using a different combination of k (number of clusters) and N (number of PCs) (Fig 1.a below). The values of S for all the models are presented in the Supplement, Figs. S1 and S2). For all models, the Silhouette coefficients for the optimal k (the higher the S, the better the performance of clustering) are better for lower N, for example, at N=3. But this N explains, in general, less than 60% of the variance.
We needed to choose N, which should include enough variability. At the same time, the performance of the clustering method using PC-applied Z500 needs to be better than using the original (non-PC-applied) Z500 fields.

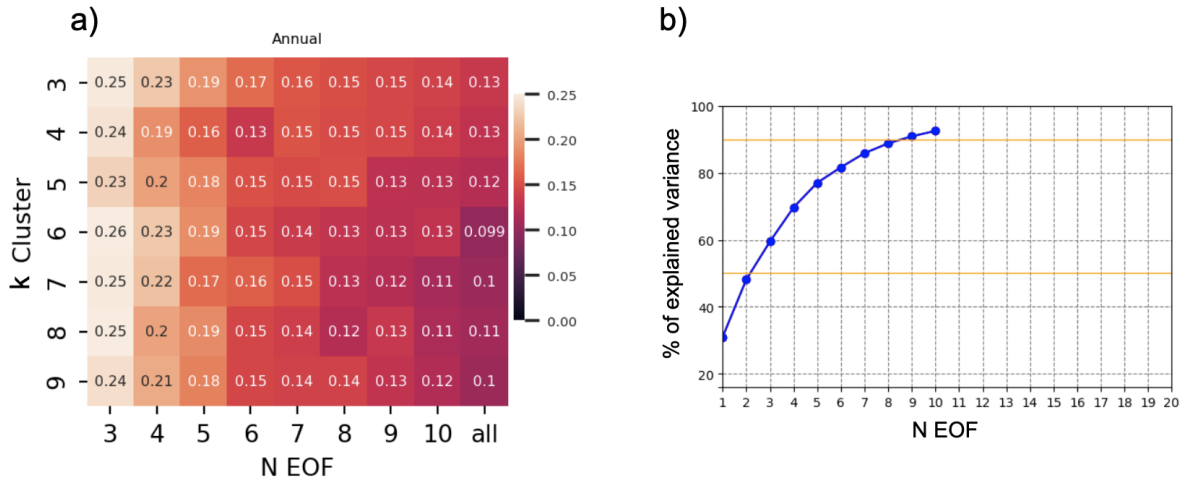


*Fig 1. a) Silhouette coefficients for a range of k clusters and N EOF for CCSM4 in Z500 during drought years in the eastern Mediterranean, and b) Percentage of explained variance with N EOF for CCSM4 in the entire Mediterranean. Yellow lines indicate 50% and 90% variance.*

This led to choosing 70% of the variance, which is the N value around 5 to 6, with k around 3 to 7.

We checked that clustered patterns do not change significantly when we use higher N or the original non-PC Z500 fields (Fig. 2) with the same number of k. Therefore, as the reviewer mentioned, the method is not sensitive to N (with N>=5, which explains at least 70% of variance) after an optimal k is chosen.

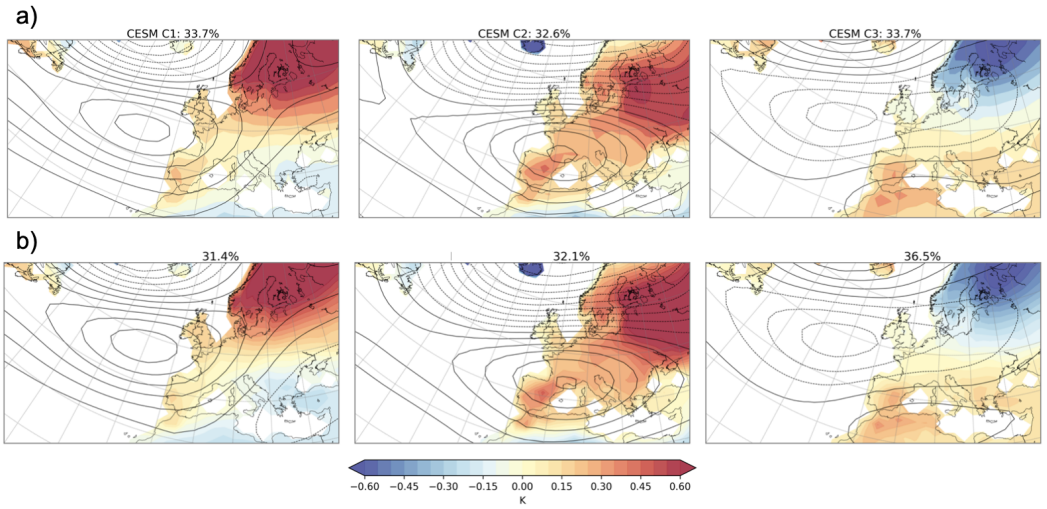


*Fig. 2. K-means clustering applied to the Z500 PC fields with (a) 7 PCs and (b) 5 PCs and with k=3 for CESM1 (Step S3 in the manuscript). Contours indicate Z500 anomalies, and colors are the temperature anomalies. Percentage values indicate percentages of years included in each cluster.*

We included in the manuscript:
[247-248] "Once an optimal k is chosen, the PC-KCA method is not sensitive to changes in N."

13) Why should N be < or = to 7?...
Thanks for the point. That is our mistake. N needs to explain more than 70% of the variance (which leads to N of 5 or 6). We deleted that part in Fig 1.

14) I understand that the dates corresponding to the identified droughts are selected and the Z500 is considered only for those dates and then the PCA is applied. Perhaps I have skipped this explanation somewhere, but I have not found it. It should be clearly explained.

We may not have been clear with it. We included this detail in:
[211] "For this, only Z500 during drought years are considered."

15) Do the xi in Eq (2) refer to the X(t,l) in eq (1)? This does not make sense to me. Can you refer the notation of Eq (2) to that of Eq (1) so that the reader can understand how the output of the PCA feeds the KCA? In my understanding these should be PC values because you

end with N modes X t drought years, with ( I assume) a PC being a time series with t time steps including drought years.

Thanks for the point. We corrected the eq (2) to be consistent with eq (1). Also, refer to our response 11).

(eq 1) $U = \sum_{i=1}^{M} \lambda_i . a_i x_i^{\top}$

(eq 2) $Q(c_1, \ldots, c_k) = \frac{1}{t} \sum_{i=1}^{t} min_{k=1,\ldots,k} || a_i - c_k ||^2$

16) Therefore, each value in a cluster could be an array of N values corresponding to how a given date (Z500 anomaly map) is represented by those N values in the space of EOFs. However this does not fit Eq (2). I may be wrong though and other approaches may be possible. What I am trying to highlight is that 3.5 needs a clear explanation of the methodological approach and its parts, with the notation of the different parts being consistent with each other.

Refer to our responses 11) and 15). We went through the method section and tried to make the section clearer.

17) The final number of clusters is 71. However, I understand that since the analysis is performed on each model, many of those will be similar. Perhaps worth commenting this here?

Thanks for the point. Each model, period (LM or Hist), and region (western or eastern Mediterranean) results in 3-7 clusters based on the similarity of the circulation patterns. We showed these numbers in Table 2, but never mentioned them in the text, so we included this detail in the revised manuscript:

[255-256] "After Step 3, 3 to 7 clusters are obtained for each model, period, and region (Table 2), totaling 71 clusters."

Then, 71 clusters are the sum of all clusters from each model. In the next step (Step 4), we performed clustering to gather these 71 clusters based on their similarities. We mentioned this in:
[260-261] "In Step 4, KCA is applied once again to these 71 clusters (from now on, referred to as *cluster*) to group similar clusters across all models, periods, and regions."

How these 71 clusters are grouped into the final 11 patterns is presented in Table A1 in the Appendix.

18) There is quite a number of typos, please revise them. Please, also in the rest of the text. I will avoid pointing at the grammar issues, but please take care of this.
We went through the manuscript to check the typos.

GC4. (*)Section4.1

19) I agree with the comments about Fig 2 and Fig 3 in general if we consider the details, maps, correlations, etc. However, perhaps I would have a different take on the actual interpretation of them. Consider Fig. 2 first. Most models have a large low-frequency variability, with large multi-decadal or multi-centennial departures from the long-term mean. Some of them often longer than the reference period considered in Fig 2. They are to a large extent not consistent among the different models, which therefore indicates that they are more obviously related to internal variability than to the external forcing common to all experiments. If this is the behaviour of a real SOIL variable then the 1950-1979 interval considered as a reference is a very short interval of time and may correspond to a very specific state in the NOAH model, assuming that it also represents reasonably reality. However, a longer integration with NOAH or if we had more observations, would supposedly show a considerable level of low frequency variability; we do not know how much because we do not know how well the models in Fig2b represent reality.

With this in mind, the fact that one simulation represents less variability during the selected period or another one represents more, does not mean that this or that model is doing better or worse, because it is not intended that these simulations represent the real 1950-79 variability, unless it would be clearly responding to external forcing or those simulations were driven by observations, which are not. The previous reasoning extends to the other arguments related to Fig3. All arguments oriented to a better or worse representation (e.g. line 324, 325), I wouldn't agree with them, because Fig 2, indicates that except for the trends in the last decades SOIL responds to internal variability and thus the maps in Fig 3 are expected to show some level of similarity but not to represent faithfully what the NOAH model does within that comparably short period.
I think the authors should consider this argument and see how it impacts the orientation of the text and the interpretation of the figure.

Following the suggestions from the reviewer, we have modified the marked sentence to avoid evaluating the worse or better representation of the relationship between Z500 and SOIL. The new sentence is now:
[326-328] "The overall comparison of pattern correlation coefficients implies that the variability of Z500 associated with SOIL in the climate models is closer to that from NOAH-LSM–ERA5 over the western region than over the eastern region."

Also, about the reviewer's comment on low-frequency variability in the models that is longer than the reference period, we modified the sentences in [334-337] to include this detail:
[334-337] "A time series of 56 years may not include all possible variability of SOIL and Z500, including low-frequency variability on multi-decadal or longer time scales present in the model simulations (Fig. 2b). This unaccounted factor could also influence the comparison between NOAH-LSM--ERA5 and climate models, hence, the significance level of the statistical tests."

Specific details:

20) It would be good to indicate in the text or in the figure caption the time interval used for correlations.
We added the time interval in the caption for Fig. 3.

[Fig. 3] "during the period 1950--2005 CE."

21) The maps shown in Fig 3 are indicating some relation to zonal circulation, NAO, which is mentioned in the text. If desired, this could be objectively calculated by indicating correlations with the NAO index in each model. But I would understand also that the authors would not want to go in that direction.

Thanks for the suggestion. As the goal of the section is to compare the Z500-soil moisture variability and spatial patterns between the observation-based dataset and the climate models and to assess their similarity during the present day rather than an understanding of involved circulations, we decided not to extend the discussion with the NAO index for this part.

GC5. Section4.2

22) 1. Figure 4 caption and in the text. The mean percentages of total drought years and the mean duration of droughts are calculated from the ensemble means...
I have reservations about the meaning of these numbers because of being calculated from the ensemble means. If the quantity that was being analysed would depend mostly on external forcing, I would agree, because the metrics based on the ensemble average would be meaningful as a filtered version obtaining after cancelling noise from internal variability. However, the behaviour of soil and drought occurrence here is shown and argued (e.g. Line 358) to be related to internal variability. Therefore, the statistics that would be comparable to the real world or representative for it according to each model are those of individual simulations, not the ensemble average. Arguing from a different angle: if we would have enough of a high number of runs, the ensemble average should tend to be flat and with no droughts.
The previous arguments also would justify why this happens (L 368): '... more clearly in those models and periods with one ensemble member.'

Thanks for the point. We included the values of mean duration and percentages of drought years for each period (Last millennium and Hist) with the standard deviations across the ensemble members in Tables A2 and A3 in the appendix.
As the reviewer commented, the standard deviations across the ensemble members indicate that there is some range of discrepancies between the ensemble members. We have included some details about it in the manuscript:

[359-363] "The percentage of drought years and the mean duration of droughts in Fig. 4 for each climate model and period, including their respective standard deviations, are presented in the appendix Tables A2 and A3. The tables show that the percentage of drought years and the mean duration of droughts vary across the ensemble members. For the percentage of drought years, particularly CESM1, CCSM4, and MIROC-ESM exhibit larger standard deviations during Hist. In the case of the mean duration, bcc-csm1-1 and MIROC-ESM show larger standard deviations during Hist over the eastern Mediterranean region."

[367-369] "The ensemble members of CESM1 and GISS-E2-R do not exhibit unanimous periods of low or high drought occurrence (figure not shown), which aligns with the

difference in the drought years and the duration of droughts across the ensemble members as presented in Tables A2 and A3."

23) 2. Line 371: ‘ This seems to agree with Cook et al (2016)’. Do you mean the simultaneous occurrence or the period or both? Please indicate more specifically and if you think it is important elaborate…

We modified the sentence to:
[384-385] "This result seems to agree with Cook et al. (2016) that have shown the simultaneous occurrence of hydroclimate variability between the western and eastern Mediterranean on a multidecadal time scale."

24) The comparison with Cook et al (2016) is interesting. For instance also in Lines 378-381. Do you think this comparison holds even if the reconstructions of Cook et al basically address the growing season (jja)? How can the statements of agreement or disagreement suffer from this? Some comments about this would be pertinent.

We repeated the wavelength coherence analysis for the annual summer soil moisture anomalies between the western and eastern Mediterranean for each model. The result is shown in Fig. 3.

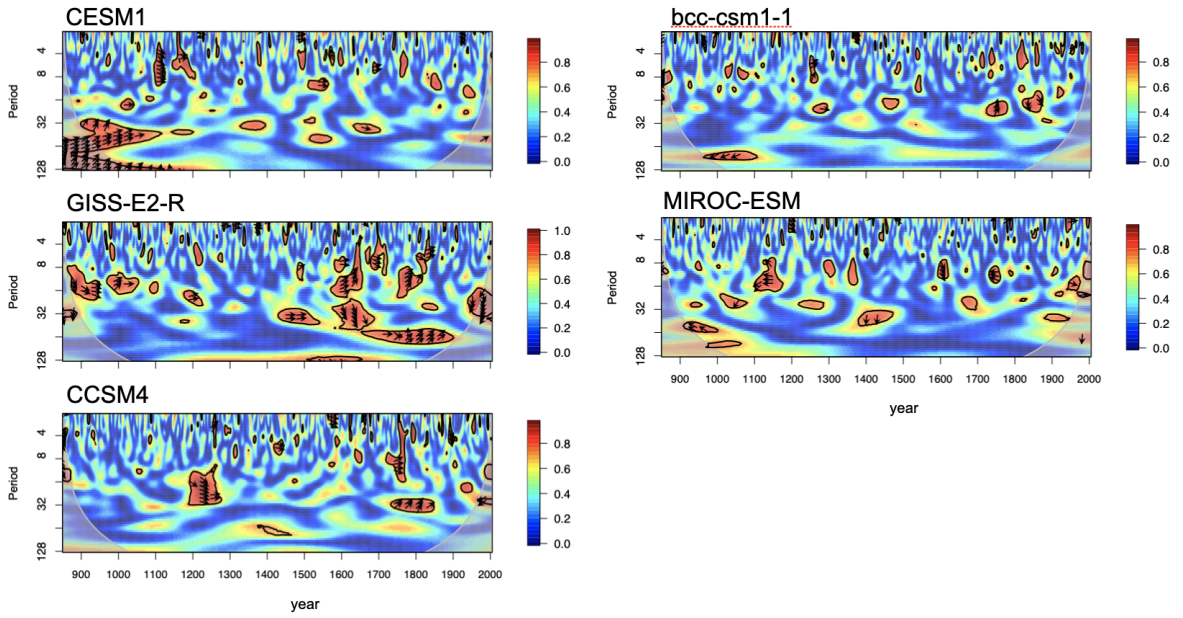


*Fig 3. Wavelength coherence analysis between the western and eastern Mediterranean for the annual summer (JJA) time series (Similar to Fig. 5a in the manuscript).*

Overall, the result is similar to those of the annual soil moisture anomalies, showing no uniform in- or out-of-phase co-variability across time-period bands. In addition, the co-variability varies depending on the model. We briefly added this detail in the revised manuscript:
[395-396] "The analysis was also repeated for the summer (JJA) SOIL (figure not shown). The summer SOIL shows the same result as the annual variability, indicating no apparent uniform phase co-variability in the climate models."

25) GC6. Line 405: ‘Although it seems contradictory that P2 depicting a negative NAO condition also occupies a significant percentage of the occurrence...
I think this could perhaps be due to the size and definition of the windows used. I do not mean that it is wrong but could be an effect of that and if so, it may be worth commenting on it. The P2 pattern favours inflow from the SW into the Iberian Peninsula. The western side of it, over Portugal and southwestern Spain should not be dry with this pattern. However, dryness could affect the lands of northern Africa and central Mediterranean Islands and over Italy. I think it is likely that the occurrence of drought with this pattern in the western box reflects the balance of wetness in the west/northwest region of the box and dryness in the rest. Perhaps it is worth assessing that and commenting.

We updated Fig. 6 with the composites of the soil moisture anomalies, and for P2, as the reviewer mentioned, we can see that the negative soil moisture anomalies are more located in the southern regions compared to the P1 patterns. We commented on this point in the revised manuscript.

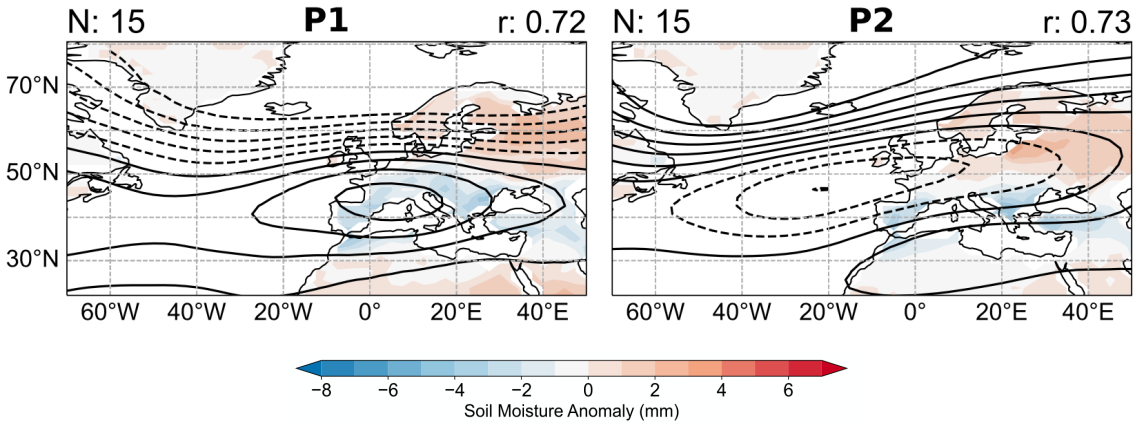


*Fig. 4. A part of the updated Fig. 6. P1 and P2 resemble a positive and negative NAO, respectively. Z500 anomalies in black contours and soil moisture anomalies in colors.*

[423-425] "Additionally, during P2, negative soil moisture anomalies associated with droughts are located predominantly in the southern Mediterranean region, indicating a higher occurrence of drought conditions in the south compared to the northern Mediterranean region. In contrast, central Europe experiences wetter conditions with negative Z500 anomalies."

26) This also takes me to suggest that it would be interesting to see the composites of soil for each group of patterns. One could actually show the composites over the whole Mediterranean, not only the boxes. This does not necessarily require an increase in the number of figures. The Z500 anomalies can be shown with lines using hatching for significance and the soil pattern with shading in the same map. It would help to understand how the different patterns influence drought in the region of interest.

We added the composites of soil moisture anomalies in Fig. 6 as suggested. Also, refer to our response 25).

[409-411] "These patterns are presented in Fig. 6 with their frequencies (in the number of occurrences per century) and the mean composites of soil moisture anomalies corresponding to each circulation pattern during 850–2005 CE."