Dr. Hugues Goosse, editor
Climate of the Past

Woon Mi Kim
Physics Institute and Oeschger Centre for Climate Change Research, University of Bern
Sidlerstrasse 5, 3012 Bern, Switzerland

16th February 2021

Dear editor Dr. Hugues Goosse,

We would like to thank you again for another opportunity to resubmit our manuscript and also to the reviewers for their detailed and insightful comments. We greatly appreciate the time that the reviewers dedicated to our manuscript. We have addressed all the reviewers' comments carefully and please find below our responses to each of the reviewers.

Sincerely,

Woon Mi Kim* and Christoph C. Raible

*Corresponding author: woonmi.kim@climate.unibe.ch

Response to the Referee 1

Minor comments

Just some minor comments on wavelet analysis: I would have suggested superimposed epoch analysis, focusing on the largest volcanic eruptions of the last millennium. If authors use wavelet, it might be helpful to give some more background information how to interpret the highly temporally irregularly distributed volcanic forcing time series in conjunction with droughts in a wavelet concept. For instance, the very large eruptions in 1258 (Samalas), 1452 (Kuwae), 1600 (Huaynaptina) and 1815 (Tambora) are reflected as clear peaks in the decadal bands for the CESM drought Indices vs. volcanic eruption wavelet coherence figures. The peak prior to 1100 AD might just be an artefact, since no large volcanic eruption in the period 850 – 1200 AD are evidenced in the reconstructed volcanic time series of GAO et al (2008) that was used as volcanic forcing for CESM. An argument for the physical implausibility of an inversion of the lead-lag droughts vs. volcanic eruptions relationship (of the pre-1100 AD peaks) is that droughts typically do not lead to large tropical eruptions.

In addition, the null hypothesis for the test is not clear to me (i.e. are the confidence intervals based on white or red noise?). At least the authors should mention those issues in an additional paragraph in the methods and respective discussion section 3.2.

Thanks for your feedback and comments. As the referee mentioned, we noted the problem with applying the wavelet coherence analysis to the time series of discrete events and we are aware of it. The process of filtering frequencies of low variability in the time series of eruptions can enlarge the effects of pre- and post-eruptions, therefore, adding some non-physical artefacts in the filtered time series.

In the revised manuscript, we still included the wavelet coherence analysis in Fig 6., as the analysis is useful to show the effects of large and small eruptions on drought indices. We now mentioned the problem associated with the wavelet coherence analysis with the time series of eruptions in lines 360 - 365, then, as suggested by the reviewer, we performed a superposed epoch analysis on the drought indices and included it in Fig. 7. The superposed epoch analysis supports a connection between the drought indices and large eruptions indicating that eruptions cause wet conditions rather than dry conditions over the Mediterranean region.

The discussion on these new results is presented in lines 350 – 373.

Lines 350 – 373: 'To further assess a potential connection between the drought indices and volcanic eruptions, a wavelet coherence analysis is applied (Fig. 6). The analysis shows that significant co-variability between the simulated drought indices (SOIL and scPDSI) and eruptions are found during periods with strong and frequent volcanic eruptions, for instance, around 1257 Samala and 1600 Huaynaputina eruptions. For small eruptions, the signals of co-variability are not uniform among the eruptions, some showing significant co-variability while other not. In addition, the phase relationships between the eruptions and the drought indices also vary among the eruptions (not shown). This non-uniformity of co-variability between the small eruptions and the drought indices is a strong indication that no physical connection between both exists, i.e., the significant co-variability is merely due to statistical artefacts. OWDA does not show a strong significant co-variability during the 1257 Samala eruption, which was the strongest eruption in the last millennium

(Gao et al., 2008). Still, OWDA shows a significant co-variability during the period of Little Ice Age around 1400 - 1600 AD, similarly to CESM.

The wavelet coherence analysis is clearly useful to distinguish the effects of strong and small eruptions on the variability of drought indices. However, the analysis poses some problems in handling discontinuous time series with a sporadic occurrence of the events, such as the volcanic eruptions. Filtering certain frequency bands from this kind of discontinuous time series smears out the eruption (i.e., an eruption starts earlier and last longer than in reality), therefore, adding some non-physical artefacts in the time series. Some of these are also reflected in Fig. 6 showing that the significant co-variability occur earlier that the actual eruption years.

Hence, for a more detailed analysis on causal effects of volcanic eruptions on drought variability, the superposed epoch analysis is applied to the 16 largest eruptions and the 16 smallest eruptions (Fig. 7; for the list of the eruption years, see table A2). The analysis shows that the increases in drought indices are followed after large eruptions, and this positive association lasts up to 3 years in CESM. On the other hand, no significant response of drought indices to small eruptions is noted. This finding is in line with Rao et al. (2017) and Mcconnell et al. (2020) that demonstrated wetter conditions in the Mediterranean region after strong eruptions.

Thus, the analysis shows that large eruptions are associated with an increase in drought indices, i.e., wet periods. In other words, the occurrence of yearly and multi-year Mediterranean droughts are not driven by the volcanic eruptions, but the internal variability.'

Response to the Referee 2

(1) The comparison of CESM with instrumental/reanalysis data: there is much improvement here (e.g., comparison of geopotential height, SST anomalies in simulations and instrumental/reanalysis), but the authors still don't show that the model is simulating the timing and magnitude of trends in local rainfall and SOIL/PDSI (e.g., Figure 12 – if the drying trend is in fact a forced signal that is as strong as the CESM appears to indicate it is, it should show up in the instrumental data)- given that the conclusions about ongoing/future forced changes hinge on the model's ability to do this, I suggest addition of instrumental vs CESM time series data in Fig 12 or a panel c in Figure 3 that shows time series from the model and observations on top of each other (I know the paper/figures are already quite long, so I don't want the authors to have to add extra figures).

We included the time series of summer scPDSI from the observation, CESM, and OWDA in Fig. 3.(c), also indicating the values of their trends on the upper left in the same figure.

We added a sentence about these trends in:

Lines 254 - 257: 'Nevertheless, all three scPDSI show negative trends during 1901-2000 AD (Fig. 3.(c)), also in each sub-period (1901 - 1950 and 1951 - 2000 AD), indicating a continuous increase of drying over the region, which is in line with the previous studies (Mariotti et al. (2008); Sousa et al. (2011); Spinoni et al. (2015)). Based on the Mann-Kendall tests, these trends are all statistically significant at 95% confidence level.'

(2 - a) Potentially flawed use of statistical tests: the authors claim to use a Mann-Whitney test to determine if the 'means' of the OWDA and CESM PDSI variables are distinguishable and find 'the means are statistically similar to one another' - this is a non-sensical test as far as I can tell. First, it looks like the two PDSI time series are normalized to similar time periods (the last millennium), so they should both be centered about a long-term mean of zero, so testing if two time series' means are statistically different if they are centered at zero is not a test that the model is doing well- the real test seems to me to be in Figure 5a (are drought distributions similar).

It is true that the time series of drought indices (scPDSI, SPI, SPEI) are normalized with respect to their long-term means. However, note that what we have compared here are the time series of duration of drought events, and not the entire time series which contain wet/dry fluctuations.

Statistically, drought episodes are located at the lower tails of the probabilistic distributions of drought indices. Hence, the time series with only droughts are clearly not centered at a long-term mean of zero and they do not necessarily follow a normal distribution. Thus, for these time series, we can use a non-parametric Mann-Whitney U test.

(2 - b) Finally, the Mann-Whitney test should be applied to test the similarity of distributions of data, not means (as far as I know), but the authors' wording suggests they are using this test to determine if means of distributions are significantly different throughout the manuscript.

We admit that we used a wrong word 'means' to describe what is tested in a Mann-Whitney U test (M-W). As far as we know, a M-W test can be interpreted as an analog to the conventional t-test for location parameters (means or medians) of two samples with unknown probabilistic distributions (Wilks, 2001). The null hypothesis of M-W test states equal distributions of two samples, and statistically equal distributions in a non-parametric test imply equal location parameters between two samples. However, as the assumption in a non-parametric test is that the location parameters of samples are still unknown, we agree that here we cannot state explicitly that the means or medians of two time series are compared to each other. Therefore, as the referee points out, we corrected the term 'means' to 'distributions' where the M-W tests are involved in our manuscript and mentioned what are tested in M-W in the method section (Sect. 2.2).

Lines 174 – 175: 'A Mann-Whitney U significance test is performed to statistically compare the distributions between the transient and control simulations.'

Wilks, Daniel S. Statistical methods in the atmospheric sciences. Vol. 100. Academic press, 2011.

(3) The reported finding that Mediterranean droughts are mainly driven by internal variability in the climate system: the wavelet coherence figure seems to demonstrate that Mediterranean PDSI shows significant coherence with the volcanic forcing time series after large eruptions, so the timing of variability doesn't look like it's only due to internal variability. Perhaps there are wet periods following volcanic eruptions, but there is not enough information in the figure caption to determine the phase of the relationship. Additionally, I would think the authors would want to show that the magnitude/duration of droughts in the control run overlap with the magnitude/duration in the forced run (and statistically test if these drought duration distributions are distinguishable), but I don't see a figure or analysis that shows this (e.g., additional box/whisker plot w CESM control drought durations in Fig. 5a)- the authors do present the mean number of droughts and durations in the control and forced runs in Section 3.3, but showing the actual drought distributions and testing this difference could be much more informative and actually test/support the authors' assertion that drought (duration) is driven by internal variability.

It is true that the wavelet coherence analysis shows some co-variability between the drought indices and eruptions, mostly after some large eruptions. However, we noted that the wavelet coherence analysis cannot fully demonstrate in detail the associations between these two variables as the usage of a discrete time series like the volcanic forcing initiates some problems. The process of filtering the frequencies of low variability in the time series of eruptions seems to enlarge the effects of pre- and post-eruptions, therefore, adding some non-physical artefacts in the filtered time series. This problem was also pointed out by the referee 1.

To show this point, we applied the low pass filters to the summer scPDSI and volcanic eruptions for some periods where the covariability between these two timeseries are significant in the wavelet

coherence map (Fig 1 below). The filtered time series of eruptions shows that the peak of the eruptions is smeared out, so that the eruptions occur much earlier than their actual years.

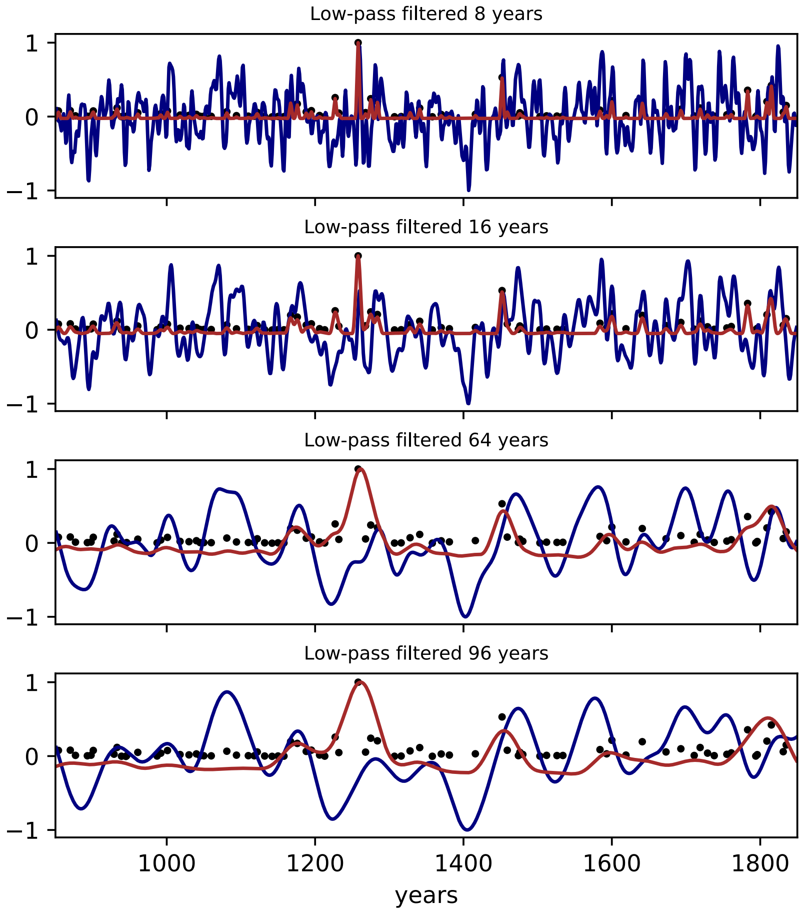

Fig 1. Low-pass filtered time series of volcanic eruptions (brown) and summer scPDSI (navy blue). The filtered time series are scaled with respect to their maximum values and standardized to make easier the visual comparison between two time series. Dots indicate the actual eruptions.

This fact is reflected in the wavelet coherence plot (Fig. 6 before the revision) that exhibits some significant signals of covariability already some years before the peaks of the eruptions. This issue also can explain why sometimes scPDSI leads the causal association with the volcanic eruptions.

However, we still found the wavelet coherence analysis useful to distinguish the effects of large and small eruptions on the variability of drought indices. Most significant covariability are found during the periods of large eruptions, while for small eruptions, the significant signals are low or random among eruptions.

Hence, we decided to still include the wavelet coherence analysis in Fig 6. but without showing the phase relationship to concentrate on discerning the effects of large and small eruptions on drought

indices. In addition, we mentioned the problem related to the wavelet coherence analysis with this specific time series in lines 360 - 365. For more robust analysis to see the association between the drought indices and volcanic eruptions, we performed the superposed epoch analysis on the drought indices. It is found that the superposed epoch analysis (Fig. 7) supports the positive association between the drought indices and large eruptions indicating that the eruptions cause wet conditions but not dry conditions. We included the plot on this analysis in Fig 7, and modified the discussion on this part in lines 350 – 373.

Also, as suggested, we included the duration of droughts from the control simulation in Fig 5. (a) and (b), and applied the M-W test to these variables between the transient and control simulations (p-values in table A1 in the appendix). Here, we noted that the variability of duration of droughts in the transient simulation is in the range of the variability found in the control simulation. The text about this result is added in lines 343 - 345.

Lines 350 - 374: 'To further assess a potential connection between the drought indices and volcanic eruptions, a wavelet coherence analysis is applied (Fig. 6). The analysis shows that significant co-variability between the simulated drought indices (SOIL and scPDSI) and eruptions are found during periods with strong and frequent volcanic eruptions, for instance, around 1257 Samala and 1600 Huaynaputina eruptions. For small eruptions, the signals of co-variability are not uniform among the eruptions, some showing significant co-variability while other not. In addition, the phase relationships between the eruptions and the drought indices also vary among the eruptions (not shown). This non-uniformity of co-variability between the small eruptions and the drought indices is a strong indication that no physical connection between both exists, i.e., the significant co-variability is merely due to statistical artefacts. OWDA does not show a strong significant co-variability during the 1257 Samala eruption, which was the strongest eruption in the last millennium (Gao et al., 2008). Still, OWDA shows a significant co-variability during the period of Little Ice Age around 1400 - 1600 AD, similarly to CESM.

The wavelet coherence analysis is clearly useful to distinguish the effects of strong and small eruptions on the variability of drought indices. However, the analysis poses some problems in handling discontinuous time series with a sporadic occurrence of the events, such as the volcanic eruptions. Filtering certain frequency bands from this kind of discontinuous time series smears out the eruption (i.e., an eruption starts earlier and last longer than in reality), therefore, adding some non-physical artefacts in the time series. Some of these are also reflected in Fig. 6 showing that the significant co-variability occur earlier that the actual eruption years.

Hence, for a more detailed analysis on causal effects of volcanic eruptions on drought variability, the superposed epoch analysis is applied to the 16 largest eruptions and the 16 smallest eruptions (Fig. 7; for the list of the eruption years, see table A2). The analysis shows that the increases in drought indices are followed after large eruptions, and this positive association lasts up to 3 years in CESM. On the other hand, no significant response of drought indices to small eruptions is noted. This finding is in line with Rao et al. (2017) and Mcconnell et al. (2020) that demonstrated wetter conditions in the Mediterranean region after strong eruptions.

Thus, the analysis shows that large eruptions are associated with an increase in drought indices, i.e., wet periods. In other words, the occurrence of yearly and multi-year Mediterranean droughts are not driven by the volcanic eruptions, but the internal variability.'

Lines 343 - 345: 'Importantly, for the same indices, the distributions of duration of droughts are statistically similar to the distributions in the control simulation at 99% confidence interval (Fig. 5. (a) and (b) and. Table A1). This implies that the variability of droughts in the transient simulation is within the range given by the internal variability in the control simulation.'

(4) The drought initiation/termination is quite interesting, but I don't know what to make of the indices (NAO/ENSO) after they have been redefined to percentiles during non-drought periods- is this standard practice? And how is this information meaningful for interpreting year to year NAO/ENSO as it relates to drought predictability/trajectories? For example, I am unsure as to what the 'extreme positive' NAO now means- how much is this index now changed if it is redefined only during non-drought years?

The motivation for this part was that the positive NAO and negative ENSO also occur frequently during the non-drought period as one can observe in Fig. 10. (a). Then our question was, how the NAO and ENSO during droughts differ in terms of their occurrences and magnitudes from those during non-droughts. For this, it was necessary to set a reference period to be a "drought-free" period without any perturbation introduced by droughts to discern clearly what happens during droughts with respect to the non-drought periods. Thus, defining the thresholds based on the non-drought period as a reference facilitates the comparison between these two periods.

Regarding the predictability/trajectories, we do not think that using these relative thresholds would lead to a different conclusion about the evolution of NAO/ENSO in different stages of droughts from using the absolute thresholds. The reason is simply that the values of the threshold for positive NAO (0.58) and negative ENSO (-0.56) (see the table 1 below) do not differ much from the absolute thresholds (0.5 and -0.5). Thus, the conclusion that the role of the large-scale circulation patterns becomes weaker through the stages of droughts remains unchanged.

The extreme NAO and ENSO are defined relative to the non-drought period with the same reason as mentioned in the first paragraph. Again, the 5th and 95th percentiles of each mode during the non-drought period are not too different from those during the total period (Table 1). These 5th and 95th extreme values differ much from those during the drought period, and this supports our approach that the condition during droughts can be characterized better by comparing with respect to the "drought-free" condition.

Regarding our approach of taking a specific period relative to another as a reference is rather common, e.g., the superposed epoch analysis to extract volcanic forcing imprints is such an example. This estimates the mean departure of climate conditions during volcanic events relative to other years.

	NAO			ENSO		
	total	**Non-drought**	**Drought**	**total**	**Non-drought**	**Drought**
5th	**-1.69**	**-1.74**	**-1.22**	**-1.59**	**-1.53**	**-1.72**
25th	**-0.64**	**-0.69**	**-0.40**	**-0.66**	**-0.56**	**-0.89**
75th	**0.69**	**0.58**	**0.98**	**0.56**	**0.61**	**0.44**
95th	**1.55**	**1.43**	**1.76**	**1.77**	**1.84**	**1.40**

Table 1. Percentiles of the distributions of NAO and ENSO

In the revised manuscript, we clarified that the thresholds we use for the analysis are relative to the non-drought period, including the extreme thresholds:

Lines 433 – 438: 'The extreme NAO and ENSO are also defined in a similar way by taking the values below 5th percentile for negative extremes and above 5th percentile for positive extremes. Defining thresholds relative to the non-drought period facilitates the comparison between the two periods, showing whether and how these modes of variability during droughts differ from those during the opposite hydroclimate conditions. For simplicity, we call these relative negative and positive phases simply as positive and negative NAO and ENSO without referring constantly that they are defined relative to the non-drought period.'

(5) I still find many sentences/phrases hard to decipher, as they are either grammatically incorrect or just hard to interpret- I have noted in line numbers where this is most apparent.

We went through the manuscript and tried to correct ambiguous sentences and also modified the parts and sentences noted by the referee.

<u>Specific comments in sections/figures/by lines:</u>

(6) Abstract: I see the authors have responded to my previous comment about model limitations and added to the main text, but again, I am left with no sense in the Abstract about any model limitations/bias – given that many readers may only look at figures and the Abstract, it would be informative to insert a phrase/sentence about model strengths and limitations. If the list of model shortcomings as it relates to simulation of drought over the Mediterranean is so long that it can't be easily summarized in the abstract, that's a problem I think.

We included some sentences on the model limitation in the last part of the abstract.

Lines 2 – 8: 'Overall, the model is able to realistically represent droughts over this region, although it shows some biases in representing El Niño Southern Oscillation (ENSO) variability and mesoscale phenomena that are relevant in the context of droughts over the region.'

Also note that we added some more information in the abstract.

(7) Lines 3-4 'Our study indicates…mainly driven by internal variability' – as I mentioned in the general comments, after seeing the wavelet coherence plots, I think this statement is flawed and needs to be qualified- perhaps the overall duration/severity of droughts is statistically indistinguishable in the control and forced simulations (which isn't shown as far as I can determine), but PDSI variability and large volcanic eruptions appear to vary coherently at interannual-decadal timescales. To really be able to make the claim that the duration/severity of drought is purely internal, the authors need to actually compare the duration/severity of control run droughts and forced simulation droughts, but I don't see this comparison anywhere. As it stands, it looks like the CESM shows a sensitivity to volcanic eruptions that the OWDA doesn't show.

We addressed this issue in the response (3).

(8) Lines 20-26: I appreciate that the authors acknowledged there are various types of drought, but I disagree with the idea that a meteorological drought that is long enough just becomes the other types of droughts. For example, couldn't an area become warmer (and receive the same amount of precipitation, experience the same drought frequency), but experience earlier spring melt, more evaporation, and thus hydrologic or agricultural drought? This ag drought would not just be caused be a persistent meteorological drought.

We agree that there are other causes of agricultural and hydrological droughts and they are not always caused by meteorological droughts. Here we wanted to state a connection among different types of droughts we mentioned, mainly when a drought is initiated by a deficit of precipitation, thus a meteorological drought. Therefore, we modified the sentence to:

Lines 30 - 34: 'If a meteorological drought lasts for a longer period, it has the potential to propagate to other types of droughts, such as agricultural or hydrological drought. In this sense, different types of droughts can become connected to each other. Thus, meteorological drought is one of the causes of other types of droughts, among other processes such as seasonal changes of run-off or an increase in evapotranspiration demand.'

(9) Lines 34-35: 'The climate of the Mediterranean is characterized as semi-arid with a pronounced annual cycle, thus, high temporal and spatial variability of the availability of water resources'- does a semi-arid climate imply spatial variability? Or temporal variability? I thought it just had to do with the mean climatologic conditions (winter wet, summer dry on average), but I could be wrong.

As the referee said, the temporal variability is related to a pronounced annual cycle. We corrected the sentence as:

Lines 43 - 44: 'The climate of the Mediterranean is characterized as semi-arid with a pronounced annual cycle, which means a high temporal variability of the availability of water resources.'

(10) Lines 38-39: 'The western and eastern regions show different precipitation regimes' this makes it sound like the region should be split into two separate regimes for the analysis.

We focus on the western-central region. The eastern region with a different precipitation regime is not included in the analysis. We mentioned our region of study and validated the choice of the region in the introduction lines 120 - 125 and in the method section (Sect. 2.2) lines 161 - 165.

(11) Line 44/49: authors define EA-WR and use it, but then ER-WR is used on line 77 (?)

Thanks for the point. That was a typo and we corrected it to Eastern Atlantic pattern (EA).

(12) Line 51: suggest 'response of the Mediterranean climate to ENSO'

We corrected the word as suggested.

(13) Line 71: 'multi-years long desiccation' is a bit of an odd phrase- suggest 'multi-year drought' or multi-year dry periods'.

We changed the word to 'multi-year dry periods'.

(14) Lines 112-113: I understand that other cited studies have shown that parts of Europe may be drying using different data sources, but it would be helpful to show the actual time series in instrumental and CESM data earlier in the manuscript, both to illustrate this visually for the reader and to validate the model's trends/drought sensitivity to warming.

We addressed this issue in the response (1).

(15) Line 132-133: suggest 'validate the model simulation used here' in place of 'our model'

We modified the sentence to 'validate the model simulation'.

(16) Line 166-167: 'The statistical tests to compare the transient to the control simulations are performed with the Mann-Whitney U significance test for the means at a 5% confidence level' – the authors go on to use this test in the text/figure captions to show that the 'means are statistically different'- similar to my general comment above, doesn't this test ask if the distributions are likely different?

We addressed this issue in (2-b) and modified the text accordingly.

(17) Lines 168-169: I'm a bit confused about which years are used and why 5 sets of 89 years are chosen for the transient simulation- the control run is 400 years, so the total number of years doesn't match in the random draw of the transient vs the control. Also, are the draws of years contiguous (e.g., years 1-89 in a row) or randomly drawn? If they are randomly drawn, what is the sense in using different sets of draws? Please clarify because this approach doesn't make immediate sense to me.

We picked 5 sets of random 89 years from the 225 drought years in the transient simulation (not in the control simulation). We performed this step to see whether the length of the simulation, therefore the years with droughts, is what forces the transient simulation to have the similar geopotential and temperature patterns to those in the control simulation during droughts (in other words, whether performing a mean over more drought years smooths some possible external signals involved during droughts). If droughts in both simulations are affected by the same drivers, then when we pick some random years with droughts from the transient simulation, we would expect to see similar circulation patterns in both transient and control simulations. In the end, the result shows that regardless the length of the simulation, the circulation patterns during droughts in both simulations are statistically similar. However, with Section 3.2 and after the response (3) to compare the duration of droughts between the transient and control simulation in the manuscript, we noted that this is a redundant analysis to affirm the same conclusion as Section 3.2. Thus, in the revised manuscript, we excluded the sentence about this step.

(18) Lines 171-174: After my last comment in the previous round of reviews, I appreciate the authors tried to split up the time periods from which they remove trends, but I still am not sure why linear trends are removed over these time periods as the figure showing the time series isn't shown until Figure 12 - it may help to show the reader these time series when discussing the time series and trend removal.

We included the figure of time series of soil moisture we used with the magnitudes of trends for each period (the 1850 – 2000 and the 2001 – 2099) in the appendix Fig. A1. and mentioned about this figure this in the line 178.

(19) Line 184: 'some drought metrics'- suggest 'various'? Or state exactly how many- 'some' sounds strange

We changed this to 'We use four drought metrics [...]'.

(20) Lines 242-243: If the model underestimates/simulates 30-50% lower precip than observations, does this bias mean anything for simulation of drought in the region? I ask because the authors mention land-atmosphere feedbacks - if the land is already 'too dry' in the model, are there implications for the life cycle of droughts (e.g., is the model somehow unrealistically 'on the edge' of setting off land-atmosphere drought feedbacks that wouldn't occur if the model just simulated a slightly wetter background climate regime that was more realistic?)

Although the mean background climate is drier than the observed value, droughts and the variables for land-atmosphere feedbacks (anomalies of temperature, geopotential height and soil moisture) are

defined with respect to the long-term mean climate. If the model showed slightly wetter climate, these droughts and land-atmosphere feedbacks would be re-defined based on the new mean condition. Thus, the result would not change much, unless the model simulates more wetter episodes that frequently deviate from the mean climate. However, in that case, these continuous wetter episodes would be reflected to a change in the annual cycle. In the end, when we compare the model with the observation (Section 3.1), we see that the difference in the mean climate between both datasets does not implicate more drought events in the model simulation (Fig 3).

(21) Line 245-246: 'the means are statistically similar to one another'- as I mention in the general comments, this appears to be a non-sensical test- PDSI should be centered around zero, so testing the means are statistically different is not a test that the model is doing well. I am also not sure if the instrumental- model-OWDA comparison was done on time series that were re-normalized to the same time period, or if the distributions from the data normalized to the last millennium are compared to distributions of data normalized to the last 100 years- if the time series are normalized to different time periods, I would expect there could be differences in the mean, etc.

We also addressed this issue in the response (2). For the last 100 years, the reference period of all datasets is the 1950 – 1979. As the time series are calibrated with respect to that 30-yr period, we do not expect them to have a mean zero over entire time period, even less when we consider only drought events.

(22) Figure 3 caption:' The p-value from the t-student test between the summer scPDSI from CESM and OWDA is 0.28' – what exactly is being tested here? I assume it's the statistical tests mentioned in the Methods, but please clarify in figure caption as I'm not sure (e.g., that distributions are significantly different or what?

Before the sentence 'The p-value […]', we added a sentence about the null hypothesis of the test: ' The t-tests are applied under the null hypothesis of an equal mean between two time series.'. As in this case, we used the t-tests and not the Mann-Whitney test (as we compare the entire time series of annual drought indices), the null hypothesis of comparing means between two sample distributions is correct.

(23) Separately, the figure caption states the 'red points on each box show the data points' – so there are only 3-5 data points? If there are <5 data points, then what is the sense in showing box plots that are intended to summarize large numbers of data points (don't the boxplots just show the minimum, first quartile, median, third quartile, and maximum? So how can 3 points of data meaningfully produce a boxplot?). Please clarify.

We agree that the boxplot is not appropriate for this case with few data points. As we added it simply for a visualization purpose, we eliminated the boxes on the data points in Fig. 3 in the revised manuscript.

(24) Also, in panel (a) the observations are black, then in (b) the colors show different information- this seems unnecessary confusing as there's no figure legend in the boxplot panel. Can the authors use consistent colors in both panels?

We changed the colors of points in the (b) to be the same as in the (a).

(25) Finally, what time periods are shown in the different parts of panel (b)? Please label on figure and/or describe in caption- for example, Figure 5a is immediately much more clear with the labels above the different sections of boxplots.

We added the labels and titles for the Fig. 3.

(26) Lines 255-256: confusing wording: 'simulating droughts of few years long, and with longer duration than those from OWDA.'

We modified the sentence to:

Lines 263 - 265: 'However, the model still shows to a certain extent its ability to reproduce persistent droughts of multi-year long duration. Additionally, the mean duration of droughts in CESM is longer than in OWDA.'

(27) Line 266-267: 'tree ring based reconstructions tend to deviate in their spectral behavior' – what does this mean?

We clarified the sentence by changing it to:

Lines 274 – 276: ' Moreover, tree ring-based reconstructions for droughts tend to overestimate low-frequency variability compared to the instrumental observations.'

(28) Line 286: confusing wording: 'despite the fact presents some discrepancy to the observation exist'

We changed the sentence to:

Lines 297 – 298: 'Although there are some discrepancies between the model simulation and the observation, the model is able to reproduce the climate conditions associated with the variability of present-day scPDSI.'

(29) Lines 286-287: 'the model does not significantly underestimate the persistence of multi-year droughts' – yet my reading of Figure 3b in which the observations are compared to the CESM show that the CESM simulates droughts that are about half the duration as compared to obs (large differences in median, inner quartile range). Perhaps the authors mean CESM simulates longer droughts than the OWDA?

Indeed, yes, it is associated with the temporal variability. Thus, we modified the sentence to:

Lines 298 – 299: 'In particular, the model is able to simulate multi-year droughts and these droughts have longer duration than those in OWDA.'

(30) Line 291: suggest 'by focusing' (not and focusing)

We changed this as suggested.

(31) Lines 293-295: 'We do not aim to make a direct comparison between the proxies and the model simulation, as this cannot be made due to the different initial conditions' – this is fair, but according to the wavelet analysis shown later, the CESM PDSI variability appears to line up with large eruptions, whereas they don't in the OWDA – if we 'believe' the model simulation, the PDSI variability could be (partly) forced, meaning they should line up temporally in the proxy data and CESM, right?

Similar comment on lines 304-307: 'are not mainly driven by external natural forcings'

We modified the analysis and interpretation of the results for this part according to (3).

The superposed epoch analysis (Fig. 7) and comparison of the transient and control simulations (Fig. 5) support that Mediterranean droughts are more associated to the internal variability than the volcanic frocings. Wet periods are followed after some large eruptions. Thus, the statement that in CESM, droughts are driven by the internal variability remains unaffected.

(32) Line 314: suggest 'statistically indistinguishable', not 'indifferent' – both here and elsewhere (I think indifferent means 'unconcerned' or 'mediocre', not indistinguishable)

Thanks for the correction. We corrected the word to 'statistically indistinguishable' or 'statistically similar'.

(33) Lines 329-331- given the timescales of impacts of volcanic forcing (e.g., multi-year cooling, then recovery to 'normal', maybe with some impacts on AMO), would you expect to see imprints of volcanic eruptions visually on drought indices that have been smoothed with a 100-year running mean?

We performed the analysis with the non-smoothed annual time series. The figure of 100-year running means is included in order to visualize better all the time series of drought indices together. We clarified the time series we used for the analysis in the caption of Fig 5.

(34) Line 334: for figure 6, the authors have shown a wavelet coherence diagram here, with no time series (and no indication which variable is the 'first' or 'second' variable as they ask the reader to interpret for lead/lag/phase relationship in the caption)- I think it would be helpful to plot the time series of volcanic eruptions below or above the plots so the readers can 'see' when the volcanic eruptions occur- for example, I know there are large eruptions ~1257/1258, ~1450, ~1600-1650, and in the early 19th century. All of these time intervals/years happen to show significant coherence with simulated scPDSI variability at ~4-16 yr timescales in figure 6, but this information is not provided for the reader, so unless the reader regularly works with the last millennium forcing data, they may not notice that there are conspicuous areas of significant coherence around large eruptions.

We added the time series of volcanic eruptions in the upper panel of Fig. 6.

(35) Lines 334-336: 'are not uniform across the period frequency bands' - Would we expect the signals of statistically significant coherent variability between drought indices to be uniform across frequency bands? I think a lack of coherence across frequency bands does not show a lack of forcing response, it only shows a lack of forcing response at certain frequencies/periods. This relationship (at longer time periods/lower frequencies) could make sense given the nature of simulated responses to volcanic eruptions in the North Atlantic in the NCAR CESM model. For example, Otto-Bliesner et al. (2016, BAMS, CESM LME documentation paper)- shows that the AMO/AMV is clearly impacted at decadal timescales by volcanic eruptions (see Figures 11 and 12 in Otto-Bliesner).

We modified all this part according to (3).

(36) Lines 339-341: 'the analysis confirms…not driven by the volcanic eruptions' – I strongly disagree with this interpretation- the figure seems to show me that the CESM PDSI, and soil variable all show 'significant' coherence with the 1258 eruption (and even some coherence with the 15th century, 17th century, and 19th century eruptions for PDSI) – the 'red' regions of coherence surrounded by black lines to show significance at ~4-16 year periods is pretty hard to miss.

The information I take away from this figure is that the OWDA does NOT show coherence with eruptions, but the CESM does, which suggests the forcing is implemented wrong in the model, or that OWDA data aren't picking up on volcanic eruptions.

We addressed this issue in the responses (3) and (31).

(37) Lines 343-345: 'The focus on one indicator is motivated by the fact…' - I still don't see any indication that the 10cm soil variable actually reflects what is happening with deeper soil water in the model- as Berg et al. 2017 (GRL) show, 10cm soil water can basically just mimic what is happening in atmospheric-centric variables likes precipitation. For example, as Berg et al. (GRL, 2017) show in their Figure 1 and describe in the text: 'In contrast, projections of negative changes in total soil moisture are more muted, in both extent and amplitude. Regions of negative changes (e.g., southern U.S. and Central America, northern South America, Mediterranean region, and South Africa) display relative changes of reduced

amplitude compared to surface changes.' – so in fact, I would argue that unless the authors show that the deeper soil moisture column actually shows the same degree of water stress, this statement is not supported by these citations.

We agree with the referee, that our statement is not supported by the citation. Thus, we modified the paragraph as:

Lines 375 - 379: 'We use SOIL as it reflects the regional hydrological balance associated with the precipitation and evapotranspiration. Another advantage of this index is that the variable is a direct output from the model, thus, it does not require any further step, except for calculating the anomalies, and statistical assumptions as other indices do. In addition, the SOIL overlaps full or a part of drought periods given by the other three indices, without significantly underestimating the multi-year duration of droughts.'

And we discuss more about this in the conclusion lines 600 - 609.

Lines 600 – 609: 'The reason is that most of the commonly used drought indices are based on a water balance that considers only the atmospheric moisture supply and demand, and these indices tend to overestimate drought risks in the future warming scenario (Berg et al., 2017, Mukherjee et al., 2018, Swann et al., 2016). Berg et al. (2017) found that surface-based indices indicate droughts, whereas the mean 3-meter soil moisture shows wet or relatively weak dry conditions compared to the surface level. In our study, we used the upper 10 cm soil moisture anomaly that partially reflects the water stress on plants. However, the upper 10 cm of soil level is not enough to fully assess the complex atmosphere-soil-vegetation interaction and the variability in the deeper levels of the soil. Thus, the upper 10 cm soil moisture used here also magnifies drought risks to some extent. However, the Mediterranean is one of the regions where still the depletion of soil moisture occurs both at the surface and in the mean 3-meter soil level, though the amplitude of the rate of decrease is reduced in the 3-meter soil moisture compared to the rate in the surface soil moisture (Berg et al., 2017).'

(38) Line 362: suggest 'presents an average of 7.25 droughts per century'- because this is an average, right?

We corrected the word as suggested.

(39) Line 404: suggest 'in the following section' (not 'in the following')

We changed the words to 'Here'.

(40) Line 407: 'The phases of NAO and ENSO are defined with respect to the non-drought periods: the values below (above 75) percentile of NAO and ENSO during the non-droughts periods are considered as negative (positive) phases of NAO and ENSO respectively (Fig. 9).' – I am unsure of what to make of this- so the authors redefined standard indices based on non-drought years? What does this do to the time series/what is the reason to do this other than to maximize drought signal of NAO and ENSO? And how is this information meaningful for 'real world' NAO or ENSO (e.g., how can this information about NAO/ENSO

during drought be used if the indices have to be redefined during non-drought years? I can see how maps of mean differences in drought and non-drought years could make sense, but I don't know how to interpret the treatment of the indices).

We addressed this issue in the response (4).

(41) Line 418: Again, is this what the MW-U test is testing (difference in means?)

We modified the MW-U test related texts according to (2).

(42) Line2 425-426: 'The positive NAO occupies 49% in the initiation years, then it decreases throughout the development of droughts, falling to 29% in the termination years' – in terms of this being meaningful information for drought prediction or giving information about how these droughts initiate, it sounds like positive NAO occurs almost exactly half of the time at the start of drought, but not traditional NAO, but instead NAO as defined by NAO during non-drought years?

Yes, 49% with respect to the relative threshold, meaning that there is more probability that the NAO with positive values (with more than 0.58 based on the table 1) will occur during droughts than during the non-drought period by 24% (as the occurrence of positive NAO during the non-drought period is always 25%, thus, 49% minus 25%). Analogously, zero to positive NAO will occur with 74% (49% + 25%) of probability during the initiation years of droughts, which reflects the shift of the distribution of NAO to positive values (as one can see in Fig. 10) compared to the distribution during the non-drought period.

In the revised manuscript, we included a sentence mentioning about this increase in the occurrence:

Lines 455 - 456: 'This shows that the occurrence of positive NAO almost doubles in the initiation years of droughts compared to the non-drought period (25% of the occurrence of positive NAO).'

(43) Line 449 vs line 457: 'transition years' vs 'transient years'- suggest consistent usage of terminology, here it makes sense (transition years), but in other locations (e.g., line 457), 'transient' is used (and in figure captions I think?)- this change in wording can be confusing, especially because the transient forcing/simulations terminology is also used, so I suggest removing wording that refers to transient as transition drought years, and just consistently use transition. (unless I completely don't understand and the authors intended there to be a difference)

Thanks for the point. That was a typo. We corrected all the 'transient' years to 'transition' years.

(44) Line 469-470: About the time series shown in Figure 12 – these would seem to suggest that the Mediterranean region is in a long-term drought relative to the last millennium- the smoothed time series for SOIL never reach pre-industrial moisture levels after ~1860 AD- does this mean that climate change has caused a long-term drought that has lasted for ~150 years? I am not a Mediterranean climate expert, so I plotted GPCCv2018 annual precipitation, as well as Dai NCAR PDSI over the relevant time periods for the Mediterranean region and see no noticeable long-term trend in precipitation, and either a drought or

a 'step function' in PDSI in the late 20th century (again, no long-term aridification trend as the CESM seems to simulate). Can the authors plot the instrumental time series in the background to show if the model exceeds the envelope of variability in the instrumental data and/or if the trends are present in instrumental data too?

We added the plot of times series of scPDSI from the meteorological station data (U.Delaware v5.01), CESM and OWDA in Fig. 3.(c) according to the comment (3). The scPDSIs here are calculated relative to 1950 – 1979 AD. In the figure, all three scPDSIs shows decreases during the last 100 years and also during last 50 years. These negative trends are statistically significant at 95% based on Mann-Kendall trend tests (null hypothesis of the test is that there is a no monotonous trend in the time series). Obviously, the magnitudes of these trends cannot be compared to the magnitudes of the trends for SOIL and precipitation in Fig. 12, which are calculated relative to the 850 years before the pre-industrial period.

Long-term decreases in precipitation and drought indices over the west and central Mediterranean region have been already reported in several studies: with scPDSI, Sousa et al. (2011), with SPEI and SPI, Spinoni et al. (2017), and with precipitation datasets, Nunes and Lourenço (2015) over Portugal, Valdes-Abellan et al. (2017) over southeastern Spain, Caloiero et al. (2018) and Philandras et al. (2011) over the entire Mediterranean region among others.

Caloiero, Tommaso, Paola Caloiero, and Francesco Frustaci. "Long-term precipitation trend analysis in Europe and in the Mediterranean basin." Water and Environment Journal 32.3 (2018): 433-445.

Nunes, A. N., and L. Lourenço. "Precipitation variability in Portugal from 1960 to 2011." Journal of Geographical Sciences 25.7 (2015): 784-800.

Philandras, C. M., et al. "Long term precipitation trends and variability within the Mediterranean region." Natural Hazards and Earth System Sciences 11.12 (2011): 3235-3250.

Sousa, Paulo M., et al. "Trends and extremes of drought indices throughout the 20th century in the Mediterranean." Natural Hazards and Earth System Sciences 11.1 (2011): 33-51.

Spinoni, Jonathan, Gustavo Naumann, and Jürgen V. Vogt. "Pan-European seasonal trends and recent changes of drought frequency and severity." Global and Planetary Change 148 (2017): 113-130.

Valdes-Abellan, Javier, M. A. Pardo, and Antonio José Tenza-Abril. "Observed precipitation trend changes in the western Mediterranean region." International Journal of Climatology 37 (2017): 1285-1296.

(45) For Figure 13, please define the acronyms ND and D (I assume this is Detrend and Non-Detrended, but this is not explicitly defined)

We modified the 'ND' to non-detrend' and 'D' to 'detrend' in the legend of Fig. 14 and also defined them in the caption.

(46) Line 504: suggest word other than ‘indifference’

As suggested in (32), we corrected this word in the manuscript.

(47) Lines 505-507: ‘this result shows that the natural mechanisms associated with droughts remain the same…’ ok, in so far as they are defined by circulation patterns, but what about increases in evapotranspiration/aridification due to increasing temperatures? This sentence basically is contradicted by the next sentence, which states that the mechanisms are anthropogenically driven- can the authors distinguish/clarify? The authors have shown how different drought drivers progress in Figure 11, but couldn’t EV change in the future, thus the droughts would not have ‘natural mechanisms’?

We agree that we were not very clear with this part. Here, our questions are how the mechanisms associated with droughts change during this period because of the anthropogenic influences and whether besides the anthropogenic effects, there are some natural effects that drive changes in droughts and drought associated mechanisms. The linear detrending was performed in order to see this point, considering that the linear trends in drought related variables are caused by the anthropogenic influences.

By the end, we confirmed that the dryness in the future is mainly due to the anthropogenic influences on the land – atmosphere feedback and not by changes related to circulation patterns.

We modified the paragraphs on introducing the motivation and discussing the result to clarify our point:

Lines 506 - 510: ‘We examine whether the mechanisms associated with Mediterranean droughts described in the previous section are affected by the anthropogenic influences on climate and whether these changes contribute to the intensification of droughts and eventual aridification in the region occurring in this period. For this, the detrending method is applied to the simulation following the steps mentioned in Sect.2.2. First, we analyze the non-detrended drought related variables with the anthropogenic influences on them; then, the detrended variables to see the background climate during droughts excluding the linear trends.’

Lines 536 - 540: ' Hence, when the linear trends, thus the anthropogenic influence, are not taken into account, the mechanisms involved in droughts remain unchanged during this period which indicates that no other factor than the anthropogenic influences in temperature is the cause of the severe dryness in 1850 - 2099 AD. In the future scenario, the intensities of both land - atmosphere feedbacks are magnified due to the increases in GP and TS caused by the increases in GHG, and these feedbacks become the dominant one at controlling the desiccation over the region.'

(48) Also, again, there is no precipitation and/or obs-based soil moisture/PDSI shown here- do the observations show the same general trends in terms of long-term aridification? If not, this an important thing to point out, if so, then great- the model is doing well, and this should be noted.

We addressed this issue in the response (44).

(49) For Figure 7 caption: 'the regions where the means between the control and transient simulations are statistically not significant at 5% confidence level'? - this seems like the wrong test here- are we interested in the means being the same, or are we interested in where the 'spread' from internal variability in the control run is different than the forced run spread?

We modified this as 'distribution' now in Fig. 8. Also refer to our response (2).

(50) Line 515: suggest changing wording to 'although our result shows' or the sentence is incomplete/comma splice

We modified the sentence as suggested.

(51) Lines 525-529: Authors conclude there is no 'causal connection between volcanic eruptions and dry conditions', but their wavelet figures indicate that volcanic eruptions are significantly coherent with PDSI and soil moisture variability at ~4-20yr periods around large eruptions. I agree that if we average all the geopotential height patterns during drought in the control and forced simulations, there may be minimal differences, but not all droughts/pluvials occur during eruptions, so any anomalous behavior after eruptions could be 'averaged out' by the large numbers of PDSI anomalies (droughts/pluvials) that do not occur after eruptions.

We addressed this issue in the response (3).

(52) Lines 553-554- These results would be much more believable if the authors showed GPCC/CRU/UDel precipitation on top of the model precipitation time series, and instrumental-based scPDSI on the CESM PDSI time series to show the model is getting the timing and magnitude of trends right: https://psl.noaa.gov/data/gridded/data.pdsi.html

We addressed this issue in the response (1).

(53) Lines 574-575: again, the authors are choosing to study a region that Berg et al. have shown has 10cm soil moisture that magnifies droughts and does not reflect what is happening in 'full column' soil moisture (to ~3m depth) – so bringing up that the authors have used 10cm soil water isn't really showing that they have got around this problem.

We agree with the referee that the statement is not supported by the citation. Thus, we corrected the paragraph as:

Lines 602 - 609: 'Berg et al. (2017) found that the surface-based indices indicate droughts, whereas the mean 3-meter soil moisture shows wet or relatively weak dry conditions compared to the surface level. In our study, we used the upper 10 cm soil moisture anomaly that partially reflects the water stress on plants. However, the upper 10 cm of soil level is not enough to fully assess the complex atmosphere-soil-vegetation interaction and the variability in the deeper levels of the soil. Thus, the upper 10 cm soil moisture used here also magnifies drought risks to some extent. However, the Mediterranean is one of

the regions where still the depletion of soil moisture occurs both at the surface and in the mean 3-meter soil level, though the amplitude of the rate of decrease is reduced in the 3-meter soil moisture compared to the rate in the surface soil moisture (Berg et al., 2017).'