Response to reviewers (second round) for manuscript cp-2020-133
Willem Huiskamp and Shayne McGregor
23/07/21

We thank both the anonymous reviewer and the editor for their helpful comments and suggestions. Below, we address each comment with the reviewers text in bold, and our responses following. Excerpts from the text are presented in italics, with changes in red.

Reviewer 1:
L.124: " For the sake of simplicity, we employ annual mean (Jan-Dec) fields for sea level pressure, surface air temperature and precipitation and focus instead on the impact of network size and calibration window length rather than seasonal effects."

I see this statement has been added to address a comment from the other reviewer (C. D.). Even though the authors were already using annual timeseries in the previous version, there are actually issues with this. The proxy network size and the calibration window length recommended at the end of the study could be affected by the fact of adding seasonal effects in proxy records. Indeed, one of the main purpose of this study is L. 80: "2) How does the geographical distribution of the proxies affect reconstruction skill?", but the effect of geographical distribution of proxies related to SAM variations is also dependent of the season targeted by these proxies (as stated by the authors in the same paragraph). Also, L. 81: "3) Are any regions in our model framework prone to producing non-stationary proxies and what could be modulating the SAM-proxy teleconnection?". Would these non-stationarities be still present for the seasonal averages of each region? How can we be sure that the non-stationarities detected in this study could not come from the correlations altered by the use of annual averages? If this study aims at better understanding how to reconstruct the SAM in the real world, what can the reader learn from the output of this model-based study if it makes conclusions from an unplausible situation in the "real world" (i.e. all proxies measuring annual averages of climate)? The authors should try to add further discussions to argue this choice because the sake of simplicity is not enough since reconstructing the SAM in the real world is a complex Problem.

We appreciate the concern raised by the reviewer regarding our selection of annual-mean climatological fields for our pseudoproxies and its relevance to real-world reconstructions. Our motivation was as follows: 1) As stated on L124, limiting our experimental parameter space and avoiding not only the issue of seasonality, but the problems that arise with this (i.e. ice cores are annually resolved - should we be combining these with seasonal pseudo-tree rings?). 2) As noted in the discussion (L405), current state of the art climate models are considerably less skillful in simulating the SAM on a seasonal time-scale than in the annual-mean. The use of seasonal data would be less valuable if the location/ intensity of the westerlies show greater biases. 3) (L405) The hope that reconstructing the SAM on an annual-mean time-scale would smooth out high-frequency noise and enhance the signal to noise ratio of our reconstructions.

This should appear in the methods rather than the discussion, and as such has been moved. The updated methods read as follows:

"With this in mind, we employ annual mean (Jan-Dec) fields for sea level pressure, surface air temperature and precipitation for the following reasons. 1) The CMIP5 generation of models (including CM2.1) are less skilful at representing seasonal variability in the SAM-SAT relationship thanover the annual-mean (Marshall and Bracegirdle, 2015); 2) Reconstructing the SAM on an annual-mean time-scale should smooth out high-frequency noise in the proxies and enhance the signal to noise ratio of our reconstructions. 3) To simplify the experimental parameter space and focus instead on the impact of network size and calibration window length rather than seasonal effects."

The reviewer's second concern is regarding the validity of these results, i.e. - given the enormous complexity of paleo-SAM reconstructions and the nuances involved in calibrating individual proxies, what value does a simplified framework such as this provide? To begin with, our method finds analogues in Abram et al. 2014 and Dätwyler et al. 2018, both of whom reconstruct an annual-mean SAM despite the inclusion of highly seasonal proxies such as tree rings (as well as annually resolved proxies such as isotopes derived from ice cores), so we disagree that this represents an entirely 'unplausible(sic) situation in the real world'. We are simply being consistent by averaging all our data annually.

Secondly, the impact of this approach as opposed to a purely seasonal approach may be hinted at in the results of Dätwyler et al. 2020, which we discuss on lines 413-417. This shows, for reconstruction skill at least, values that are similar to those we find in our results.

L. 150: "It should be noted that as a 95% confidence interval is used, non-stationarity will be falsely identified 5% of the time, hence we define a grid point as non-stationary only if the running correlation falls out of the confidence interval more than 10% of the time, or 50 of the 500 years; more than double the 5% we might expect by chance alone."

According to the author's scripts, maps from Fig. 6 are drawn at each 0.1 step for the number of times the running correlations falls out of the ones drawn by Monte-Carlo repetitions. So, if I well understand, this means that it is not 50 of the 500 years but rather 10% of the 471, 441 and 411 sliding time frames tested for the 91, 61 and 31 length cases (respectively). If I am right, the authors should avoid saying this is equivalent to 50 of the 500 years, same as for the Fig. 6 caption. More generally, saying that a given year out of 500 has a running correlation doesn't make sense, unless if taken as a centre of a time frame, but there are not 500 of them here.

We thank the reviewer for spotting this error! They are correct that, rather than a time-frame of 50 years, the 10% threshold should be 47, 44 and 41 years for the 31, 61 and 91 running correlation windows, respectively. As such, Figure 6 in the manuscript has been updated and the caption now reads as follows:

"Figure 6. Number of years at each grid point where the 31 year (a,b), 61 year (c,d) and 91 year (e,f) running correlation between SAT (a,c,e) or precipitation (b,d,f) and the model SAM falls outside the 95% 'stationarity' confidence interval (Section 2.2). As per our definition of non-stationarity, regions which fall outside this interval 10% of the time or more (≥47, 44 and 41 years for the 31, 61 and 91 windows, respectively) are highlighted with solid black contours and are considered to be non-stationary"

L. 154: "As correlations are bounded between +/-1, the running correlations are converted to Fisher Z-scores:"
Yes, correlations are bounded by +/-1, so? I don't see why this is stated here while not a single Z-score is discussed later. Are they used to compute significance levels within the Monte-Carlo framework? If yes, it should be said somewhere because in the present form, it feels these Z-scores come from nowhere.

We thank the reviewer for noticing this - it was not written clearly enough. They are indeed correct that the Z-scores are used to calculate the significance levels within the Monte-Carlo framework. This sentence has been updated and now reads as follows:

"The running correlations are converted to Fisher Z-scores to ensure they are normally distributed for the calculation of confidence intervals"

L. 192: This is not clear why a band-pass filter is applied to n3.4. How has the latter been chosen? Would this significantly change the correlations significances calculated by the authors if not using the band-pass filtering?

The band-pass filter is applied to the model nino3.4 index due to its exclusive comparison with time-averaged values of proxy-SAM correlations. By calculating these running means between (for example) precipitation and the model SAM, we are smoothing out variability shorter than 31, 61 or 91 years. To ensure the power spectrum of the model ENSO record was consistent, they were band pass filtered before being correlated with the SAM-proxy running correlation records. We would expect these correlations to change if we had not performed the smoothing, but the results would be spurious.

The choice of the Nino3.4 index is due to its ability to optimally represent the character and evolution of El Nino and La Nina events (Bamston et al., 1997; Trenberth and Stepaniak 2001). These references have been added to the methods section.

L. 188-196: Some uses of "correlated" are a bit confusing in those lines since it is describing how to determine which proxy/SAM teleconnections are effectively correlated with ENSO variations. It would be clearer for the reader to simply say that a correlation coefficient is computed instead.

We agree that this could have been more clearly articulated. As a result, this sentence has been replaced and now reads as follows:

"*To investigate the role ENSO may play in modulating the pseudoproxy-SAM teleconnection, a correlation coefficient is calculated between running correlation time-series' of SAM-SAT/precipitation and the model Nino3.4 (n3.4) index at each grid point.*"

L. 250: The paradigm of the choice between the calibration window length and the amount of available proxy data fully covering this time frame is very well discussed here. It is very challenging and strongly affects reconstructions when working with real world data. However, I would have thought that a reconstruction method such as the CPS is not so much affected by this problem because correlations (weights) can be computed for each individual overlap periods of the proxies and the SAM (which can then be maximised for each). This is different for methods like the PCR (principal component regression) for which it is impossible to diagonalise the proxy matrix if it has missing data, which thus makes important the choice of the length of the calibration period common to all proxies.

The reviewer is correct in their observation that a weighted CPS does not explicitly require the calibration period to be uniform across all proxy records, unlike methods such as PCR. It does however complicate the weighting of these proxies consistently. For example, a proxy calibrated over 91 years will likely have a lower correlation coefficient than another proxy calibrated over 31 years resulting in a reduced weighting in the final reconstruction, despite being better constrained. There is clearly no ideal solution here, but in this instance, we believe being consistent in our application of calibration window length at least eliminates this added complexity.

L. 275: Similarly to my comment for L. 150, there is something wrong here. If the authors are effectively considering running correlations for 31-, 61- and 91-time length, falling out of the 95% range in more than 10% of the time would means 48 times out of 471 for 31-year windows (because there are 471-time windows of size 31 in a 500 year-long period). In the same way, falling out of the 95% range in more than 10% of the time would means 45 times out of 441 for 61-year windows and 42 times out of 411 for the 91 ones. In Fig. 6 it is apparent that the authors use contours at level 50 (i.e., when colours turn to orange), and not at the true 10% levels, specific to each time length used to compute the running correlations (see above). This means that the authors are not rejecting stationarity when 10% of running correlations falls out of the 95% Monte-Carlo range. They are actually doing it for ~10.6% (50/471) for 31-year time length, the ~11.3% (50/441) for 61-year time length, and ~12.2% for 91-year time length. The authors need to find a way to fix Fig. 6 and subsequent analyses if they want to keep the rejection of stationarity at a 10% level. Otherwise, they should adapt the main text to this fact.

As with the previous comment for L154, these percentages have now been calculated, so that each accurately represents a true 10% of the total time interval.

L. 277: "For SAT, 4% (31 year window) and 8% (61 and 91 year window) of land cells are

non-stationary, while for precipitation 6% (31 year window, 10% (61 year window), and 9% are (91 year window).” These values might not be the same after addressing my last comment. There is also a missing right bracket in this sentence.

The reviewer is correct. These percentages have been recalculated and the changes read as follows:
“For SAT, 6% (31 year window) and 11% (61 and 91 year windows) of land cells are non-stationary, while for precipitation 7% (31 year window) and 14% are (61 and 91 year windows)”

L. 277: “[…] despite ENSO potentially being responsible for 50% of this variance.” 50% when ENSO is band-passed filtered? How has it been calculated?

As is stated in the methods, all calculations pertaining to the model ENSO index are those that have been band-pass filtered. We agree that the 50% figure was not clearly attributed to regression analysis displayed in Figure 11. A new reference has been included in this sentence as well as a slight re-wording. It now reads:

“*This is of lesser consequence, however, as most sites show little variance in SAM-SAT teleconnection at this longer window length (Figure 8c; most regions have an rstd<0.1) despite ENSO potentially being responsible for 50% or more of this variance (Figure 11c)*”

L. 433: Remove “TEXT”.

This has been removed.

Editor’s comments

Title of section 1.1. Is it reconstructions of ‘SAM variance’ or of ‘SAM variations’ ? ‘variance’ may be misleading because of the association with the statistical definition of the variance.

We agree with the editor that this could be clearer. As such, we have changed the title from ‘variance’ to ‘variations’.

Lines 83-90. Maybe add some references on the interest and limitations of pseudo-proxies?

We thank the editor for the suggestion. We have included references to the papers by Mann and Rutherford (2002) and Mann et al. (2007), which introduces the concept of pseudoproxies and explores the robustness of proxy-based reconstructions of climatic fields, respectively.

Line 85. I would use ‘mask’ instead of ‘degenerate’ and I must admit that it was not easy for me to make the link between the first half of the sentence that mentions ‘non-climatic

noise' and the second half focused on a 'real' change in the strength of teleconnection due to variability in the climate. I think I finally got the message but I had to read several times the sentence so maybe a rephrasing would be helpful

We agree that this was not sufficiently clear. The sentence has been split and rephrased, and is hopefully now unambiguous. The new excerpt reads as follows:

"These 'perfect' proxies are free from non-climatic noise that may degrade a teleconnection signal between a real proxy and the SAM. Instead, our pseudoproxies isolate changes in teleconnection strength due to underlying variability in the climate only."

Line 108, it is not clear what is referred to in 'in this instance'. Is it Karpechko et al. (2009)? Another issue for Figure 1 is the comparison of the results of control simulations with the observations over past decades that likely includes a forced trend. It is mentioned elsewhere but I think the impact of this point should be included in the discussion of the figure 1.

We have amended this sentence to add clarity and include the editors suggestion with regards to the comparison of a time period over which a trend is observed, with a pre-industrial, stable control simulation. The new sentence reads as follows:

"*As previously noted, we should be cautious directly comparing observations spanning a brief time period (in this instance, ERA-Interim (Dee et al., 2011) data correlated with the Marshall SAM index (Marshall, 2003) over the 36 year period from 1979-2014) with a well observed SAM trend, with our model data which represents a stable pre-industrial climate spanning 500 years*"

Line 168. If I understand well, the proxies are selected if they meet the criteria on one of the windows, not all of them, and you are only using one time window for all the proxies at the same time for each reconstruction. Maybe it would help to state this explicitly. The wording 'For each grid point of the model' may suggest that the procedure is different for each point.

The editor has indeed interpreted our methodology correctly. We agree that this section could have been more clearly structured. Three paragraphs in total have been altered and we have explicitly included the following sentence to address the issue of calibration over one or more windows:
"*Similarly, all sites in a network are selected based upon correlations over a single window, and may therefore be absent from networks calibrated using a different window.*"

Line 215 (and also 253-260). The larger skill when using precipitation record is interesting. Is it possible to make a link with the correlation between those records (precipitation and temperature) with SAM, as for instance illustrated in Figure 2.

We link the correlation between these records and SAM and reconstruction skill with the inclusion of Figure 8, which depicts that precipitation sites that meet our selection criteria, while fewer in number, show reduced variation in correlation with SAM over our 10 calibration windows, when compared with SAT. This is briefly addressed in the discussion (L365). This may however be one of many contributing factors (as was discussed with C.D. in the previous review round) including the distribution of sites selected for a network and how effective this is at cancelling out regional noise (this may be different for precipitation and SAT). We leave such analyses for future work.

Figure 10. I may have missed it but the reference to figure 10 occurs after the one to figure 11 in the text.

We thank the editor for spotting this error. The two figures have swapped places to be consistent with when they are cited.

Line 355. 'less exhibiting less spread', maybe suppress one 'less'.

This has been corrected.

Line 368 My visual interpretation is that the difference in RMSE is small between precipitation and temperature derived reconstructions. This difference should be quantified here to confirm that it 'again perform better'.

The editor is indeed correct that the median RMSE measure shows a far smaller difference in reconstruction skill between SAT and precipitation. That being said, precipitation-derived reconstructions still perform better overall. We have amended this sentence to reflect this, it now reads:

"If we consider median root mean square error (RMSE), precipitation-derived reconstructions perform better overall (minimum RMSE of 0.91 for SAT and 0.90 for precipitation; Figure A2). As with our threshold skill score, the RMSE shows skill is maximised by utilising a large proxy network and a longer calibration window of 61 or 91 years, though the difference in skill between SAT and precipitation is smaller."