Review of Huiskamp and McGregor

1) General comment:

I appreciate that most of the comments from my previous report have been addressed by the authors. The new version reads well, and the methodology is easier to follow. The authors often explain well to what extent their study can be useful for "real-world" SAM reconstructions and what are the limitations, which is very appreciated. I agree for the publication of this article that would be a very nice support for other pseudo-proxy experiments and real-world reconstruction of SAM variability. Although the Results section is clear and interesting, I still have concerns about a few technical aspects. There is notably something wrong, or unclear at best, with the non-stationarity test based on the one from Gallant et al. (2013).
I sincerely apologize for the time delay of my second report.

2) Revisions/Comments:

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.

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.

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.

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?

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.

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.

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.

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.

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

L. 433: Remove "TEXT".