Response to Reviewer 1

We thank the reviewer for their constructive feedback and thoughtful suggestions. Below, we address each comment with the reviewers text in bold, and our responses following. Excerpts from the text are presented in italics and additions are in red.

2.0. Abstract

• The authors should mention the complete acronyme of the model, i.e. GFDL CM2.1, at least in the abstract. More generally, the authors go back and forth between "CM2.1" and "GFDL CM2.1" in the main text. Two options:

1. They should mention it once in the main text followed by brackets mentioning the short appellation : "[...] GFDL CM2.1 (CM2.1 hereafter) [...].
2. Otherwise they should call it with the long name in the whole text and figures. Since the model is called "CM2.1" in figures, option 1 seems the less constraining.

Thank you for spotting this. We will note the model's full name in the Abstract and follow the convention of 'option 1' for the remainder of the manuscript.

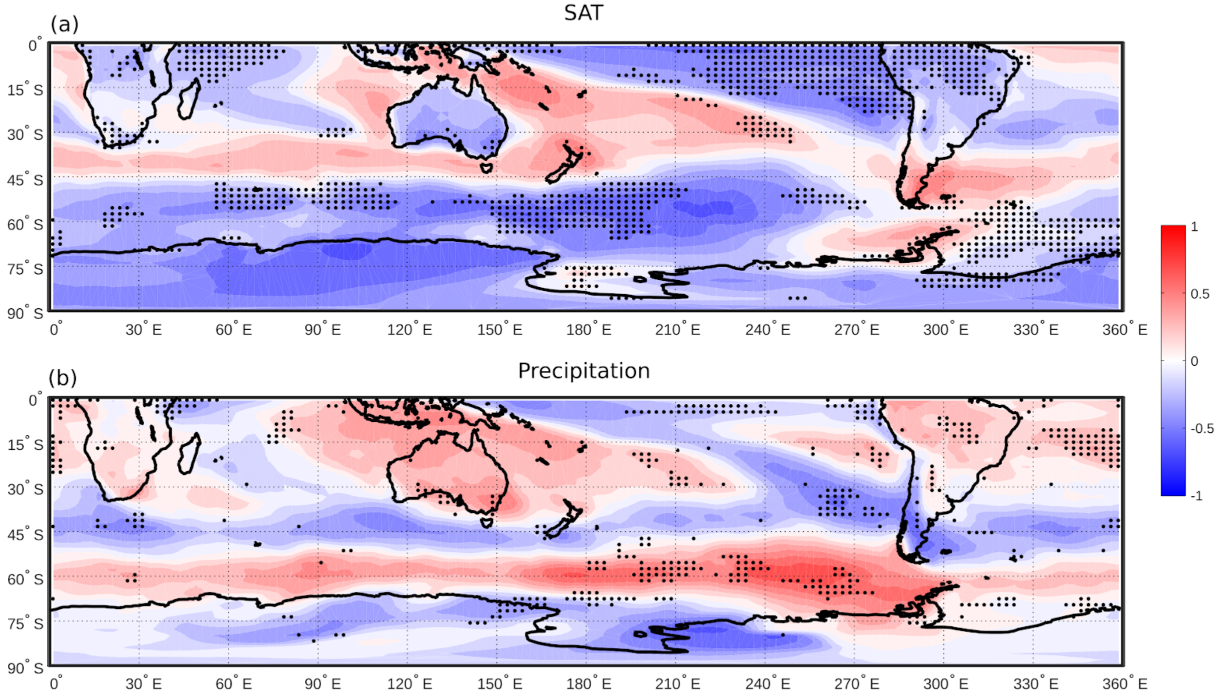
2.1. Introduction

• Figure 1: Maps are only partially informative since no confidence levels for correlations are shown. Instead, the authors use contours for |r| > 0.3, and I guess the other levels of contours are drawn every 0.1 steps (should have been mentioned in caption if that's the case). I understand these contours are useful when using continuous colorscales, I personally don't like it but that's not an issue here. However, the authors should at least add on maps which correlations are significant, and which are not (at the 90% confidence level, for instance). Indeed, for model simulations and instrumental data used here, time frames over which correlations are calculated have very different lengths (500 and 36 respectively), such that 0.3 correlation (for example) yields very different levels of significance for the two cases. For instance, the Student test from McCarthy et al. (2015) with corrected degrees of freedom can be used (see their Methods section): https://www.nature.com/articles/nature14491. I understand the authors are not showing significance levels to facilitate the comparison between observed and simulated patterns of SAM, which are indeed well matching, but the significance levels might hide this. The authors should at least address 1), and i suggest that using 2) might give a better figure, but i let the latter up to the authors:

1. The authors should keep colors and just indicate (for example) with little crosses the grid points for which correlations are not significant and explain that the 36 years time frame is too short to capture statistical significance, so we still see that GFDL CM2.1 seems adequate for testing SAM reconstructions. (See attached Fig. 1: same as Fig. 1b from the manuscript but with little black dots indicating correlations unilaterally not significant at the 90% confidence level, using Student test from McCarthy et al. 2015, see link in above paragraph).

2. The historical significance levels can probably be "boosted" using longer temperature and precipitation products, that will enable to use 21 more common years with SAM observations (1958 to 1978) and give more robust estimations of the SAM precipitation/temperature pattern. There is, for instance, the CRU TS dataset (Harris et al. 2020, https://www.nature.com/articles/s41597-020-0453-3).

We appreciate the reviewers thoughtful suggestions for improving this figure, and this was a criticism shared by reviewer two. We have decided to replace Figure 1 with a new figure that, rather than comparing the correlations over 500 years vs. 36 years, identifies whether or not the values calculated for the ERA-Interim/Marshall SAM data fall within the range of the 36 year running correlations in the model data. It can be found below:



Correlations of annual-mean (Jan-Dec) SAT (a) and precipitation (b) from the GFDL CM2.1 model with the model-derived SAM,calculated over 500 years. Black dots show where the correlation of the ERA-Interim reanalysis product with the Marshall SAM index,calculated over a 36 year period from 1979-2014, does not fall within the range of the model's 36 year running correlation at each grid cell.

• L.32: Typo. Space between "cycle" and ".".

This has been corrected.

2.2. Methods

• L89-91: This is one of the major limitations of the study I think. The authors are working on a control run while using a correlation-based weighting technique (weighted CPS). In

the real world, the observed SAM has a trend, which is likely to be due to human activity as mentioned by the authors in the introduction. Thus, since most of the temperature signals also has trends in the real world, one can expect calculating correlations (and then weights here) overestimated for the proxies that has the closer trend to the SAM one. In other words, If this study aims at being a support for real-world SAM reconstructions, the authors should mention as a caveat the fact that contrary to the pseudo proxy experiment, the correlations calculated in a real-world experiment over calibration periods might be artificially overestimated/underestimated because of the specific SAM trend.

We strongly agree with the reviewer on this point and have made additions to the discussion to emphasise this. While this does represent a caveat, our results show that even in a ‘stable’ climate, significant uncertainty in SAM reconstructions can occur, and we would expect this to simply be exacerbated by calibration over the modern, anthropogenic trend.

The new discussion paragraph reads as follows:

“A caveat of this study is our use of annual mean data, rather than seasonal fields. This is a distinction from previous real-world reconstructions utilising tree ring records (Zhang et al., 2010; Villalba et al., 2012; Abram et al., 2014; Dätwyler et al., 2018) which are not only more sensitive to SAT or precipitation of a particular season, but also combine these with other proxies such as ice cores (Zhang et al., 2010; Abram et al., 2014; Dätwyler et al., 2018), corals (Zhang et al., 2010) and lake sediments (Abram et al., 2014; Dätwyler et al., 2018) each of which may be more or less seasonally sensitive to multiple climatological fields. In addition, many proxies such as tree rings (Cullen and Grierson, 2009; Villalba et al., 2012) have been shown to have a lag relationship with SAM from the previous year, which is also not accounted for in this study. Dätwyler et al.(2020)’s ‘perfect’ pseudoproxy experiments for an austral summer SAM show similar reconstruction skill to our results (an average 31yr running correlation of~0.7-0.8 for their ensemble mean) and while the methods of this study are not analogous to theirs, it supports the conclusion that proxy-derived reconstructions of the SAM in a model framework can, at best, reproduce 50-60% of SAM variance on an annual time-scale. Finally, the results we present here are derived from a control simulation and the uncertainties in our reconstructions represent noise internal to the climate system. This is in direct contrast to real-world reconstructions which have the bad fortune of requiring proxies to be calibrated over a period with a significant anthropologically forced trend in the SAM. We would expect this to increase the uncertainty in reconstructions and any future model-based verification of real-world reconstructions would need to address this problem.”

• L95: "CM2.1 was selected due to its relatively good representation of the SAM.". Since this is the major reason for which the authors use this model, we need an objective statistical test that shows that spatial correlations of SAM with temperature and precipitation simulated from CM2.1 are significantly similar to those from the observations at a given significance level. This is far more convincing than just "CM2.1 has a relatively good representation" with a subjective use of "relatively". I guess a

Monte-Carlo approach using distance between pairs of random correlation maps can do the job. I have no doubts that the statistical test won't affect the model choice of the authors.

We believe that the inclusion of the new Figure 1 sufficiently addresses this point, as well as the addition of reference to the Bracegirdle et al. 2020 study, which highlights CM2.1's ability to simulate the position and strength of the southern hemisphere jet favourably when compared to the CMIP5 and CMIP6 suite of models. This is of course in addition to the references already included in the manuscript which have performed far more thorough analyses on the model's performance. We have revised/rephrased the methods section, and it now reads as follows:

"CM2.1 is selected due to its good representation of the SAM compared to similar models from the CMIP5 and CMIP6 archives (Bracegirdle et al., 2020) while Karpechko et al. (2009) find performance to be favourable when compared to ERA-40data. The spatial structure of the SAM is well simulated, accurately capturing the centre of action over the Pacific, while being slightly too zonally symmetric on the eastern half of the Southern Hemisphere (Raphael and Holland, 2006). Importantly for our purposes, CM2.1 accurately simulates the latitude at which the SAM transitions from its positive to its negative phase (as expressed via regression onto 850hPa winds) over South America, which many models of a similar age and computational complexity fail to achieve (Raphael and Holland (2006), their Figure 4b). The amplitude of the model SAM index is comparable with observations (Raphael and Holland, 2006), although its variability is larger than observed (Karpechko et al., 2009). As previously noted, we should be cautious directly comparing observations (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) spanning a brief time period with our model data which spans 500 years. To address this, we calculate a 36 year running correlation between the model SAM and our SAT and precipitation fields and identify if the correlations derived from observations fall within the model range (Figure 1). The SAM index in the model is calculated according to the method of Gallant et al. (2013) as the difference in normalized, zonally averaged sea level pressure anomalies between 40◦S and 60◦S. Aside from a region in equatorial South America in the SAT field, the agreement is good, with 87% of SAT and 95% of precipitation grid cells on land showing agreement with observations."

• L144: I guess the authors mean "a_0 and a_1 are regression coefficients [...]", instead of "$a_0 + a_1$ are regression coefficients [...]"

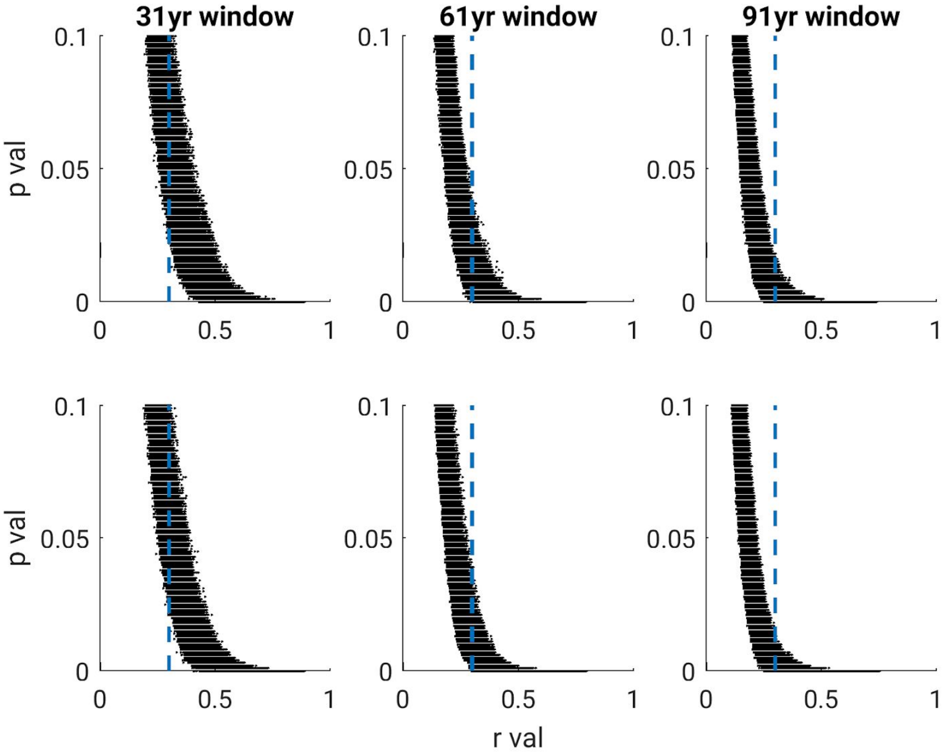
The reviewer is correct. This has been amended in the revised manuscript.

• L167: I finally understand the relevance of contours from Fig. 1. As for the comment in the introduction, the choose of |r| > 0.3 is here purely arbitrary. Would we obtain the same results with other arbitrary choices like 0.25 or 0.4? I think a way for the authors to gain in transparency in their methodological set-up is to make a selection of proxies where the correlation significance at a given level is used for selecting them, instead of using the absolute values of correlations, that have different meanings for 31, 61 and

91 window length. I know the authors have been honest and claimed that this choice is arbitrary, but I am wondering if a (pesudo)proxy selection based on a statistical test is also geographically too restrictive. If not, I think the authors should consider this way for selecting (pesudo)proxies.

We thank the reviewer for their considered comments on this issue. Our choice to use a correlation criteria finds precedence in McGregor et al. 2013 (doi:10.5194/cp-9-2269-2013) and Batehup et al. 2015 (doi:10.5194/cp-11-1733-2015) and the motivation is twofold: firstly, to set a reasonable standard for proxy skill to ensure they actually capture the SAM signal to some degree. Secondly, it highlights the result we present in Figure 2: that a ‘good’ correlation (of 0.3, for example) over a shorter window may not represent the proxies’ true correlation over 500 years and, when using a method such as an *r*-weighted CPS, this assumption should result in far more uncertainty in the range of reconstruction skill for these shorter windows. This would not be as clearly visible for some significance criteria, as a longer window results in poorer correlation coefficients meeting more stringent significance criteria (see below figure).

We would also add that the choice of 0.3 (as opposed to 0.25 or 0.4 as suggested) as a cutoff value would be similar to the choice required for a significance criteria, in other words do we select the cutoff at p < 0.1, 0.05 or 0.01? Each would similarly change the number of proxies available for selection and by extension, the resulting reconstructions.



This figure shows the running correlation values for every land cell plotted against the significance of the correlation (calculated using the Ebisuzaki method outlined in Abram et al. 2014). Panels on the top row are for SAT, while those on the bottom are for precipitation.

We have updated the methods section to include references presented above. It now reads:

"Proxies are randomly selected in accordance with several conditions. The proxy must be on land in the Southern Hemisphere and must have a correlation with the model SAM index of |*0.3*| *or greater within the calibration window after the method of McGregor et al. (2013) and Batehup et al. (2015). While a correlation of 0.3 is arbitrary in choice, it ensures that the proxy represents the SAM to some extent while not being so high that proxies are only sourced from a geographically limited region."*

In addition the discussion and conclusions section has been updated to read:

"*Increasing the calibration window does not increase the chance of producing a more skilful reconstruction, it does however,along with maximising the number of proxies, cause the range of reconstruction skill to converge on the skill of our 'perfectly' calibrated proxy reconstructions (blue envelopes, Figures 4 and 5). It should be noted that this will be, in part, due to our correlation requirement of r >=* |*0.3*| *for proxies imposing a progressively more rigorous selection criteria for longer calibration windows. Adding more sites to a reconstruction has limited benefit in terms of the maximum skill it can achieve, with values largely plateauing at a network size of~20. When it comes to minimum skill, however, this improves for increases in network sizeall the way up to and including network sizes of 70 proxies (Figure 3). This in turn, acts to increase the proportion of skilful reconstructions for a given window size.*"

• L188: This statement is way too strong. Probably that in VonStorch's 2009 study, weighted CPS is better in skill than the PCR method. However that might not be true for other datasets (maybe for the author's dataset who knows). There is indeed a very famous mathematical theorem, named the "No free lunch" theorem, that stipulates that given two statistical methods A and B, we can always find datasets for which A is better than B, and others for which B is better than A (Wolpert 1996, https://www.mitpressjournals.org/doi/abs/10.1162/neco.1996.8.7.1341). With this in mind, the authors should remove this statement and just say that the reconstruction sensitivity to the method is beyond the scope of the study.

We concur with the reviewer - this statement is indeed too strong, and has been removed as requested.

• L190-191: Are the authors calculating the skill of the correlation using the 500 years of the control simulation? If yes, this is problematic and it should be fixed, because this means that skill scores for longer calibration time frames will be artificially increased, only by the fact that they are including more data over which the statistical model is built. If not, this should be clarified in the main text because this point is not clear to me.

The reviewer is correct in their assessment - the reconstructions are validated over the full 500 years of the data (including the calibration period) as is done in Batehup *et al.* 2015. The reviewer is correct that, for real world reconstructions, we would validate over a separate period to the calibration window to ascertain an unbiased estimate of the reconstruction skill that then presumably represents its skill over the full time-interval of the data. In this instance, we have no interest in this as we can simply validate each reconstruction over the full 500 year time-period (and therefore also allowing a direct comparison to our 'perfect' or 'true' 500 year reconstructions, which by their nature require validation over their calibration period).

The effect of this, rather than artificially increasing the skill of the reconstruction, is to increase its convergence with the 'true', 500 year calibrated reconstructions we create (visible in Figures 4 and 5). For example, we can imagine reconstructions calibrated over a very long window of 470 years, then validated over only 30 years, the resulting spread of calculated 'skill' would be larger than in actuality.

We agree with the reviewer that this was not stated clearly enough in the methods section and have amended the manuscript as follows:

"To quantify the skill of the pseudoproxy reconstructions, Pearson correlation coefficients are calculated between each SAT/precipitation-derived SAM index and the sea level pressure-derived SAM index over the full 500 years of data."

2.3. Results

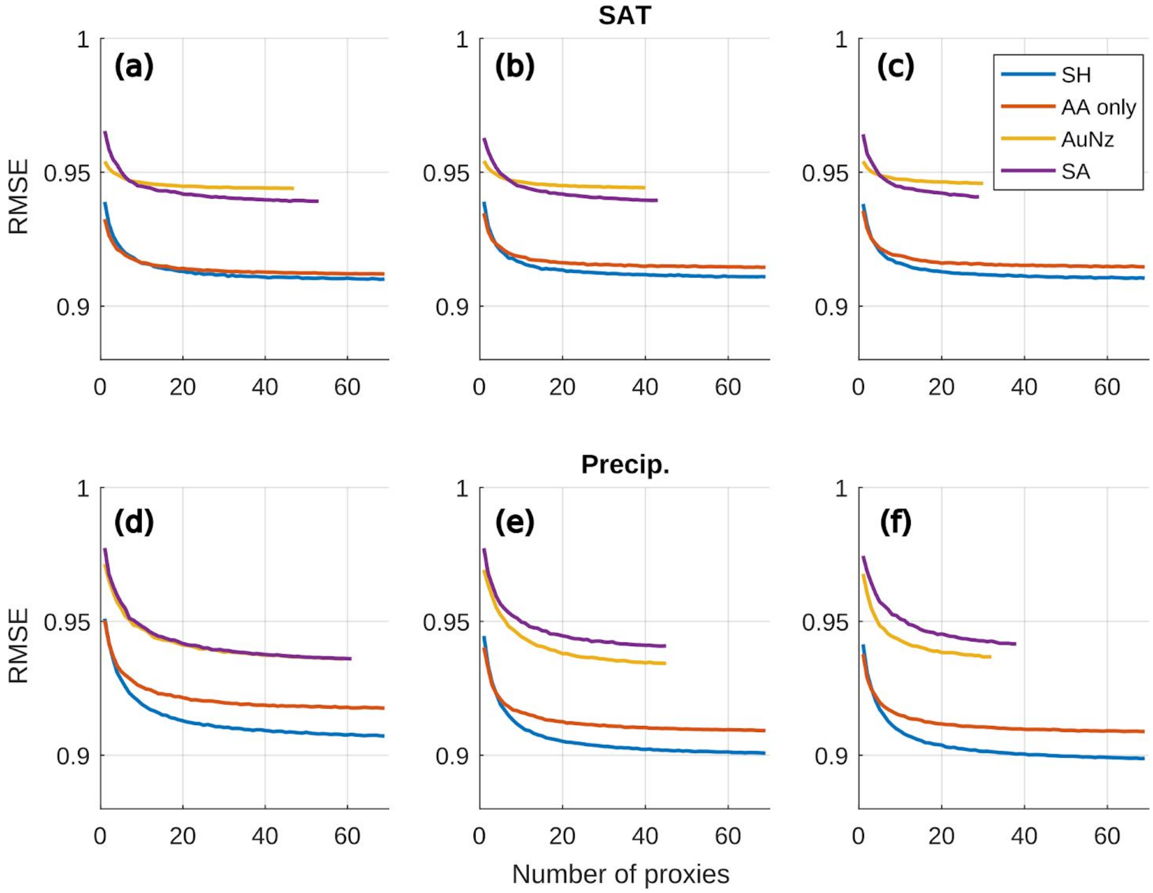
• L198: Please address my last statement from the Methods section because this might affect the first statement of this section.

The results in Figure 2 are independent of the concern the reviewer raises in their comment regarding the validation of reconstructions over the full 500 years of data. This figure merely displays, for all land points, the 'apparent' vs. 'true' correlation for a pseudoproxy.

• L211-212: Seems a bit arbitrary too to say that a reconstruction is skillful when r 2 > 0.5, although this is a relatively constraining criterion. I do not require changes here, but the authors should keep in mind that using more relevant evaluation metrics such as the coefficient of efficiency or the reduction of error (discussed for instance in Macias-Fauria et al. 2012, https://www.glaciology.net/pdf/Macias-Fauria-Dendrochronologia12-P ersistence.pdf) can clearly help for addressing this issue (and probably the last one too).

We thank the reviewer for raising this concern (shared by reviewer 2). We believe our arbitrary 'skill' criteria to be a relatively generous one, selected primarily as a way to summarise the thousands of reconstructions across our parameter space using some cutoff criteria (in this instance - can the reconstruction capture at least half the variance of the true model SAM?). We acknowledge, however, that a secondary measure would complement our approach. To that end, we will include the following figure in the Appendix. It displays the median root mean square error across the 10,000 reconstructions calculated for each network size. An additional point of discussion has been added to the manuscript which reads:

"Assessing the skilfulness of our reconstructions, where skilfulness is defined as being able to reproduce ≥ 50% of SAM variance over the full 500 years, reconstructions derived from precipitation performed best (Figure 3), with a maximum of 91% of reconstructions being reported as skilful (91 year window, 70 proxies) as well as less spread due to variability of the teleconnection between precipitation and the SAM. SAT-derived reconstructions performed poorly by comparison, with only 25% qualifying as skilful (61 year window, 70 proxies). It is worth noting that this result remains consistent when examining a different measure for skill. If we consider median root mean square error (RMSE), precipitation derived reconstructions perform better and aswith our threshold skill score, the RMSE improves as we increase the number of proxies in a reconstruction (Figure A2)."



This figure shows median RMSE across the 10,000 reconstructions calculated for each network size for SAT (a,b,c) and precipitation (d,e,f). Panels a) and d) show results for the 31yr calibration window while panels b) and e) show results for the 61yr window and c) and f) show errors for the 91yr window.

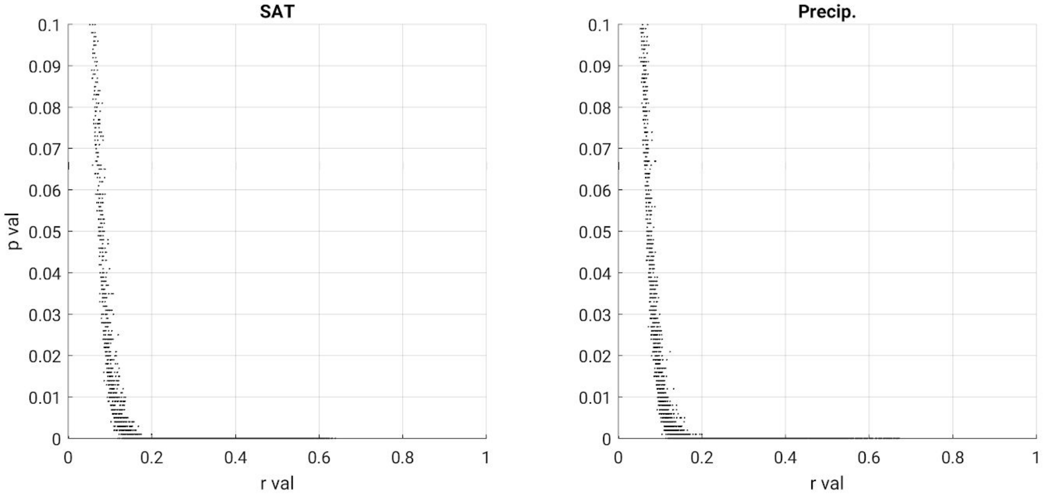
- **L230: Is there any reference for this so-called "perfect" reconstruction approach for calculating skills? I feel like this is an "handmade" skill metric, but please let me know if I**

am wrong. If I am wrong, the authors should add a reference then. Naively, i would say that the "perfect" reconstruction has 1 correlation with the pressure-based model SAM. Here, the authors' definition of perfect is purely relative to the time period covered by the control simulation. If we add 200 more years of simulation, is the 700-years based pseudoproxy reconstruction "more perfect" than the 500-years one? Considering the last two comments about the metric used and correlation calculated over the 500 years (including calibration), I think that I see why the authors use this "perfect reconstruction" skill metric. Here again, I think using more relevant metrics than correlation could avoid overfitting, handmade metrics, and some arbitrary choices.

We apologise to the reviewer that our approach was not sufficiently clear here. When we refer to a 'perfect' reconstruction, it is not one that has an r = 1 with the model SAM, but rather one where the proxies are calibrated over the entire reconstruction period (500 years in this instance). This is not intended to be a measure of skill, but rather to act as a point of comparison, displaying the limits of our reconstruction approach in the model framework we have selected. To avoid this confusion, we have changed the use of the term 'perfect' to a more appropriate 'true' reconstruction throughout the manuscript.

• L284: It's good to know that non-stationary relationships tend to fall in regions for which correlations are weak. It would have been even more interesting to see that these correlations are not significant (cf. my comments on Fig 1).

As the reviewer suggests, we have calculated the significance of the 500 yr correlations using the Ebisuzaki method described above (see figure below). We find that despite these weak correlation coefficients, virtually all are significant at p < 0.1 (to be specific, all correlations greater than 0.08). Below we present a figure visualising this. In addition, we have noted this in the manuscript as suggested. This sentence now reads:
"is also noteworthy that these non-stationary regions (as calculated using the 31 year running correlation) appear to fall, on average, in regions where correlations are weaker (though still significant at p< 0.1 when r>0.08) over the full 500 years (Figure 1a and b). "



• Figure 10: The authors say hatched points are for p > 0.5. I guess they mean p < 0.05. Otherwise it has no sense.

We thank the reviewer for spotting this mistake. It has been corrected.

2.4. Discussions and conclusions

• First two paragraphs: The authors claim that reconstructions derived from temperatures are more skillful. But with real world data, we often use both precipitation and temperature proxies. Is there any reason for not considering the case of a mixing of both types of data? Should at least been mentioned in the discussion.

A third set of reconstructions utilising both types of proxies was not pursued as the parameter space being covered in this study is already considerable. While this issue of differences with real-world reconstructions was addressed to an extent later in the discussion, we have made an addition to the following paragraph (new text in read) to more explicitly highlight this.

"A caveat of this study is our use of annual mean data, rather than seasonal fields. This is a distinction from previous real-world reconstructions utilising tree ring records (Zhang et al., 2010; Villalba et al., 2012; Abram et al., 2014; Dätwyler et al.,2018) which are not only more sensitive to SAT or precipitation of a particular season, but also combine these with other proxies such as ice cores (Zhang et al., 2010; Abram et al., 2014; Dätwyler et al., 2018), corals (Zhang et al., 2010) and lake sediments (Abram et al., 2014; Dätwyler et al., 2018) each of which may be more or less seasonally sensitive to multiple climatological fields."

• L343-344: Why comparing the proportion of skillful reconstructions of SAT and precip for 61-years and 91-years calibration window lengths, respectively?

What we report here are the best results for the respective fields (SAT and precip) - for SAT, a calibration window of 61 yrs with 70 proxies produced the best result while precipitation-derived reconstructions produced the best result when calibrated over 91 years and with 70 proxies. We apologise to the reviewer that this was not more clearly articulated and have reformulated the text. It now reads:

"Assessing the skilfulness of our reconstructions, where skilfulness is defined as being able to reproduce ≥ 50% of SAM variance over the full 500 years, reconstructions derived from precipitation performed best (Figure 3), with a maximum of 91% of reconstructions being reported as skilful (91 year window, 70 proxies) as well as exhibiting less spread due to variability of the teleconnection between precipitation and the SAM. SAT-derived reconstructions performed poorly by comparison, with only a maximum of 25% of reconstructions qualifying as skilful (61 year window, 70 proxies)."

- **L374: 1) Or maybe that 31 years windows are simply too short to robustly estimate the significance of the correlation with ENSO.**

- **L374: Suggestions 1, 2 and 3 for avoiding an ENSO bias in the reconstruction should be discussed a bit more. Do they mean that we need to wait 91 years of direct SAM observations (at least year 2049 then) that overlaps with at least 70 SH proxies (so, far more than 2049) that all are not significantly correlated with ENSO but all significantly correlated with SAM? That's personally what I concluded from these suggestions.**

We respond to the above comments together, as this section has been re-written.
Instead of making concrete statements regarding the requirements for a reconstruction free from ENSO bias, the aim here was to highlight that, similar to non-stationary proxies, we can improve our chances of a more skillful reconstruction by calibrating over a longer window. This is because the variance in proxy-SAM correlation decreases, meaning that even though ENSO may be responsible for a large proportion of this variance, it has less of an impact on the reconstruction. We have re-written this section and it now reads as follows:

"*It is not unreasonable to suspect that ENSO may be contributing towards the proxy-SAM teleconnection variance. Dätwyler et al. (2020) identify a highly variable, but centennial-average of -0.3 between austral summer ENSO and SAM reconstructions over the last millennium. Their pseudoproxy experiments using a CESM1 ensemble show significant changes in SAT during periods of large negative SAM-ENSO correlation (their Figure 4, bottom left panel). The pattern is similar to our results (Figure 9a), with regions of significant correlation over much of Antarctica and three regions in the Southern ocean centered on roughly 60°E, 150°E and 60°W. Rather than excluding proxies whose teleconnection with SAM is significantly correlated with ENSO, we can minimise ENSO's impact simply by calibrating over a longer window thus ensuring that, while ENSO may impact these proxies, the variance of their teleconnection with SAM will be small. Its greater impact at longer windows (Figure 11) is therefore minimised as the variance of the proxy-SAM teleconnection is smaller (Figure 8).*"