Clim. Past Discuss.,
https://doi.org/10.5194/cp-2020-133-RC1, 2020

Climate of the Past Discussions

Interactive comment on "Quantifying paleo-reconstruction skill of the Southern Annular Mode in a model framework" *by* Willem Huiskamp and Shayne McGregor

Anonymous Referee #1

Received and published: 18 November 2020

article [utf8]inputenc hyperref

Review of Huiskamp and McGregor

November 2020

1 Overview

This study is a pseudo-proxy experiment, based on the GFDL CM2.1 model, which explores the experimental sensitivity of hemispheric-wide reconstructions of the SAM index. I have personally tried to perform reconstructions of the historical SAM index in a "real-world" experiment and I was not convinced that it's currently feasible because of several limitations (trend in SAM, too few *in-situ* observations, too few continental proxies in SH,...). I thus think that the present study is very relevant within this context, since reconstructing SAM is very challenging. Indeed, I am convinced that the use of GCM(s) is essential for addressing the limitations in real-world reconstructions of SAM. I appreciate the way the authors have tried assessing different sources of uncertainty for the reconstruction such as non-stationarity in SAM-proxies relationships, the length of the calibration window and the size of the (pseudo-)proxy network. I also appreciate that the manuscript is well-written and generally easy to follow with a clear story-line.

However, there are some aspects in this study, that have poor statistical meaning. I would like that the authors try to address the different comments/suggestions below. Although I propose an "Accepted manuscript with minor revisions", these revisions might need running again some codes and slightly modifying some figures. However I am pretty sure that the suggestions I make won't bring large modifications in authors conclusions/discussions.

2 Comments/Suggestions

2.0. Abstract

- The authors should mention the complete acronyme of the model, *i.e.* GFDL CM2.1, at leat 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.

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).
- L.32: Typo. Space between "cycle" and ".".

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.
- 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 rep-

resentation" 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.

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

2.3. Results

- L198: Please address my last statement from the Methods section because this might affect the first statement of this section.
- 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).
- 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.
- 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).
- Figure 10: The authors say hatched points are for $p > 0.5$. I guess they mean $p < 0.05$. Otherwise it has no sense.

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