Review of the discussion paper «Quantifying paleo-reconstruction skill of the Southern Annular Mode in a model framework» submitted by W. Huiskamp and S. McGregor to the journal Climate of the Past

Christoph Dätwyler, 24 November 2020

Summary. In the submitted study, the authors use pseudo-proxy records generated from the GFDL CM2.1 climate model to assess the effect that the number of contributing records as well as their geographic distribution has on the skill of SAM reconstructions. For the pseudo-proxy-generation the surface air temperature and precipitation fields in the model are used. Furthermore, they analyse the (non-)stationarity of their pseudo-proxy records and aim at identifying the influence of ENSO on the relationship between the pseudo-proxies and the SAM index.
General comment. In my opinion, the manuscript has the potential to become a helpful contribution toward our understanding on, e.g., how non-stationarities in proxy records can affect palaeoclimate reconstructions of SAM or how sensitive SAM reconstruction are to the influence of ENSO and also other related questions. However, at the current state I have a hard time to see the scientific value the proposed study offers, because the study heavily suffers from several major issues that all need to be addressed before a publication in Climate of the Past should be considered. I try to give hints below how these issues could be addressed.
I am aware of the fact the for any manuscript every reader most likely has his/her own wish-list of things that could be done in addition to what is already presented. So in the major points below I restrict myself to what I consider a conditio sine qua non and do not ask for things that I think would be "nice to have", but are not of crucial importance.
I very much encourage the authors to take the effort to address the raised concerns and revise their manuscript, since the general topics the study touches are of major scientific importance and with the submitted manuscript already a first step has been taken toward a relevant contribution to our understanding of questions related to these topics.
Recommendation. I recommend the manuscript to be re-evaluated after major revisions.

Major points

1 **Robustness.** I have major concerns regarding the robustness of the presented results. I do have the impression that the results depend too much on several choices the authors made. Generally speaking, the results must/should not depend too heavily on such choices to be of any significant value. Or if there is a strong dependence on a choice, this choice must be very well justified. Creating a supplementary material file to complement the manuscript would allow the authors to incorporate the results of robustness tests. Before the manuscript should be considered for publication, checking the robustness of the following points is irremissible.

1.1 How strongly are the results model-dependent? It is well-known that climate models struggle to capture high-latitude dynamics. The model they use is already about 15 years old and I wonder whether there have not been any advances since then. Using a model that is as good as possible is of vital importance for this study if any of the conclusions drawn should have a meaning in real world scenarios.
Moreover, the authors also (partly) justify their choice of CM2.1 with Figure 1 and state that there is a spatially good agreement between the correlations (of SAM with SAT and precipitation) in the model and the reanalysis data (L95-97). I don't agree with this statement because if we, e.g., look at SAT and exclude Antarctica then I'd estimate (visually) that about half of the grid points that are over land (i.e. South America, Australasia and Africa) do not even share the same sign. Adding data from

one or more additional model, would strongly improve the quality of the paper, as it would allow to assess all conclusions with regards to the choice of model data.

1.2 How much different would the results look if for the pseudo-proxy selection, a different absolute value for the correlation with the model-based SAM index would be chosen? I don't even think fixing an absolute value does make sense at all because different window width for calibration are used. It is much easier to get a high correlation when correlating only over 31 years as compared to 91 years. Rather, the choice whether a record is goes into the reconstruction should be based on whether the correlation with the target is significant or not.

1.3 How much does the choice of $r^2 \geq 0.5$ / $r \geq 0.71$ when defining a "skilful reconstruction" affect the outcome? This choice is very cumbersome to me and I cannot see any justification for it. By just looking at the figures, I suspect choosing e.g. $r \geq 0.6$ or $r \geq 0.8$ would completely invalidate the conclusions that stand in connection with this measure of skill. Given that the authors screen their proxy location using correlations, it is to be expected that a reconstruction rated with a correlation-based skill score will perform rather well. Instead, a skill measure that is different from the condition used for screening would allow a more robust assessment. Furthermore, since this definition of skilful is purely looking at the correlation of the reconstruction with the model-derived SAM index, the reconstruction could theoretically be completely off and still have perfect skill, or the reconstruction could have lost almost all its variance and be nearly completely flat while still having an extremely high correlation.
There are many possibilities that could potentially help here. I am think of accuracy measures such as e.g. RMSE, RE and CE. I hope this suggestion helps the authors to find a justifiable way of measuring skill.

1.4 How strongly do the results depend on the 10%-choice in defining non-stationarity? My hunch is that this choice might be a bit less critical than the other three above, but I still suggest checking it.

2 **Language and structure.** In general the English per se is on a good level. I acknowledge that every author has his/her own style of writing and way of expressing himself/herself. However, I still feel like the whole manuscript would gain a lot by paying attention to details in formulations and also if a native speaker could read through the whole manuscript in a very detailed manner to iron out circuitously and strangely formulated sentences. Time and again I came across sentences that were inaccurate and where I had a strong feeling that a native speaker would phrase it differently. Usually I understood (I think/hope) what the authors intended to say, but it didn't read smoothly.
While in some cases the style/formulation of the content certainly is debatable and also partly a matter of taste, there are so many inconsistencies and glitches in the manuscript that at some point I stopped listing them a) because I don't want the review to be longer than the manuscript itself and it should be the responsibility of the authors to read through the manuscript *before* submission and b) because in my opinion major revisions are required that demand changing most of the text anyway.
It should not be necessary to mention, but before submitting a manuscript to a journal, the author team should take care to ensure that it meets reasonable quality standards – which regrettably was not the case here. I'm not referring to the scientific content, but to the text and figures that contain reams of inconsistencies and mistakes like for example missing and erroneous axis labels, missing panel labels ("(a)"-"(f)" in Figure 2), inconsistent font size (Figure 1), inconsistent spelling of words (skilful – skillful, grid point – gridpoint, nonstationary – non-stationary), inconsistency with abbreviations (e.g., the authors use Southern Hemisphere several time before introducing the abbreviation and when they introduce the abbreviation they do it four (!) times but thereafter keep using the spelled out

version) etc. etc.
Please also carefully select and structure the content of the different chapters. E.g., in the methods section there is a whole paragraph I have the impression does not belong there (L107-127), or Figure 1 shows up as a reference/result in the introduction where the work of others is reviewed, or as a further example, the definition of what the authors call a "skilful reconstruction" clearly belongs to the methods section.
For my taste, it is a rather long manuscript with many figures (11 in total). I have the impression it could be streamlined and some of the less relevant figures and content moved to a supplementary material file that goes along with the manuscript. As an example I think the whole page 9 could be condensed to 2-3 sentences, moved to a supplement or removed completely. The content presented here is neither novel nor surprising/unexpected but simply statistically obvious, well-known and does not require any sort of "analysis".

3 **Content.**

3.1 Similar to the example of page 9 I just made above, the whole manuscript should be streamlined and condensed to present only the essential parts which will then make enough space to address the following concerns and include the suggestions.

3.2 The authors claim in the abstract L7-8 "Non-stationarity of proxy-SAM teleconnections, as defined here, plays a negligible role in reconstructions, …", they say on L288 "To better illustrate the impact of non-stationary proxies on reconstructions, Figure 7 shows …" and also in the Discussion/Conclusion chapter (L336-338) it reads "In this study we … examine … whether or not non-stationarities in these proxy networks significantly impact the reconstructions." These statements are simply not true. I can't find any place where the authors show or analyse how these non-stationarities actually affect reconstructions. But these left-out analyses would exactly be the sort of results that would very much help making the study more valuable. E.g., I would move Fig. 7 to a supplement. What would be of interest here is how relevant non-stationarities in proxy records are for reconstructions. What is the relation to skill and how do reconstructions with a high number of non-stationary proxy records look like in comparison to reconstructions where the number non-stationary records is much lower? What is the proportion of non-stationary proxy records where non-stationarities actually become problematic for reconstructions? Just providing the chance with which a certain proportion of non-stationary SAM records will go into the reconstruction does not say anything about the effect non-stationary records have on the resulting reconstruction and its skill.

3.3 The authors sample pseudo-proxy records from random locations on the Southern Hemisphere's land mass. The results obtained with these random pseudo-proxies do not have much relevance for statements/claims/recommendations they wish to make for SAM reconstructions based on real palaeoclimate proxy records. To increase the study's relevance, I see no way around to also use the locations where we have real-world proxy records. Otherwise this study is at risk to become a pure exercise in statistics who's results are only valid and relevant for the very specific climate model that was chosen and does not provide insights that could be transferred into a broader context.
Including analyses with locations of real-world proxies will also help making what is describe on L234-235 a valid attempt, provided that the model(s) used is(are) a good enough representation of reality at these locations. At the moment I don't believe that the analyses in which Antarctica is excluded have much informative value for tree-ring-only reconstructions because a) the proxy locations are chosen randomly (on land) where mostly no real-world proxy records are available and b) for large parts the model struggles to even get the sign of the correlation right (cf. point 1.1 above and Figure 1 in the manuscript)

3.4 In the Discussion and Conclusions section (L373-377) the authors claim that the three listed points (which are very obvious and not really helpful) reduce the extent to which ENSO impacts their reconstructions. However, as under point 3.2, they don't show anywhere in the manuscript whether at all and if so to what extent these points affect their reconstructions. Analysing this possible impact is what would be of interested here. For this I suggest e.g. comparing reconstructions with randomly sampled proxy records with reconstructions that only include "good" records (according to points 1)-3)). This would then allow the authors to make a statement in the direction they aim for here, but again with the caveat that the results may only be valid for the specific model in case the agreement of the model with reality is not sufficiently good.
Also, I wonder why the stated points should be different for seasonal data (L376-377). I think also for seasonal data it is obvious that proxy records with a significant correlation to ENSO can negatively impact the reliability of the resulting reconstructions or that proxy records with a strong enough correlation to SAM should be used.

3.5 A further point I wonder about is how much the results relating to whether SAT or precipitation record-based reconstructions perform better (in terms of the currently used skill measure) go beyond what can simply be expected statistically. The authors select records based on their correlation with SAM over the calibration period. Then a skill measure that is solely based on the correlation of the reconstruction with SAM is used. Wouldn't the result whether SAT- or precipitation-based SAM reconstructions have higher skill then simply depend on the distribution of the correlations of SAM with SAT/precipitation in the model, that is, wouldn't the reconstructions based on the climate variable with a higher proportion of correlations (absolute values) that are bigger or equal than 0.3 necessarily yield higher skill values? Or in other words, if in the model you have a higher proportion of correlation above 0.3 of SAT or precipitation with SAM, then it would be much more likely by random selection to catch pseudoproxy records that correlate strongly with SAM and hence much more likely to obtain a "skilful" reconstruction.
In conclusion, the results would exclusively depend on how the correlations in the model between SAT/precipitation and SAM are distributed (not spatially) and you don't even need to do reconstructions to know the outcome/answer.

Minor points

4 L7: I think it is not necessary to introduce the abbreviation SAT here (it is done on L73).

5 L7-9: Unclear sentence. "range in reconstruction skill": Does that range in any sense relate to non-stationarity? Where is that range obtained from? If it relates to non-stationarity the sentence does not make sense, because in the first part of it you say that non-stationarity plays a negligible role in reconstructions (where skill also belongs to).

6 L13: I would remove "nominally" here.

7 Figure 1 and caption:

7.1 GFDL CM2.1 and ERA not defined yet (and I think it is not done anywhere else).

7.2 Please be consistent in how you refer to the panels (a)-(d). Here for example you use (a), (c), b), d), b and d.

7.3 Very minor detail: Consider switching (a) and (c) with (b) and (d) so that in the text there reference to (a) and (c) comes before (b) and (d).

7.4 One colour bar and one x-axis label would be sufficient I think.

7.5 Different font sizes on the colour bar are used (1 and 0 is significantly larger than -1, -0.5 and 0.5). Same for the y-axes (S from 30°S on is larger than the rest). In addition,

the number next to the colour bar are extremely close to it. Add some space, because one is tempted to read -0.5 on the red half of the colour bar.
Please pay attention to such details!

7.6 Why are contours at $r \geq |0.3|$ of relevance here? This seems arbitrary.

7.7 I don't think a the reference for ERA-Interim is needed here (you provide it on L97).

8 L15: "… have been linked …": Here results from the literature are presented, but your own Figure 1 is added as reference. You should clearly separate your own results from other people's work. I think I understand the intention here (which I guess was to refer to Figure 1 as being in line with results from the literature) but this needs to be rephrased (and I'm not sure I would include your own result in the introduction here, but this is debatable).

9 L22-24: Ok, but feels like a bit out of place and irrelevant here. I would remove this.

10 L26-27: If I remember correctly there is also some work by Dave Thompson that might be relevant here and could possibly be cited. Please check and add if you feel like (and if you don't want to include it, it could at least be interesting to read :)).

11 L28: "as derived" sounds strange to me.

12 L29: "long-term" missing before "context"?

13 L29: Unclear. Do you mean the past five decades by "present day"? Please be specific.

14 L29-30: I think the structure of this sentence does not work. I have the impression that by "present day changes" you mean the same as by "observed multi-decadal trends". Maybe just place a dot after "variability" (or otherwise rephrase).

15 L30: "… this, this …". Not very elegant…

16 L30-32: Does all this really follow from the previous sentence?? I'd remove this sentence.

17 L35: I don't think "meanwhile" is the appropriate word here.

18 L39: "can be made by examining changes" sounds strange / weird formulation…

19 L41: "… sensitive to *both* precipitation *or* surface air temperature …" you can't use "both" and then "or".

20 L42: Why is SAT not define here but only later (L73 or so)?

21 L42: "For example …". I feel like this sentence does not belong here. You already discussed this before.

22 L49: Is particle dust really relevant here and for SAM? Seems to be out of place. I'd remove this.

23 L55: Not sure about the formating here. Think I would write "… (CPS; Abram et al. …)".

24 L62: No comma after "regional".

25 L64: "comparing" instead of "compare".

26 L74: "*a* significant positive correlation" or "significant positive correlations".

27 L78-81: Isn't this true also for longer calibration windows?

28 L82: Do you really "quantify" the "uncertainties" mentioned?

29 I noticed a mix of past and present tense in the methods section. I think generally only one should be chosen and then used consistently.

30 L89: Using "is" here sounds strange ("The data … is …").

31 L94: ERA-40 does not appear in Figure 1b and d, but the way you refer to Figure 1 it should.

32 L97: You miss saying that the reanalysis data are correlated with the Marshall SAM index.

33 L100: What do you mean by "transitions from its positive to its negative node".

34 L104-105: Yes, but the present study is about annual reconstructions and later on in Section 4 it is mentioned that CMIP5 generation models have issues in representing SAM-SAT relationships on a seasonal scale. Hence, saying that Gallant et al. also use CM2.1 due to its seasonal skill (austral winter) is a very incomprehensible argument for using CM2.1 (and why is "austral winter" capitalised?)

35 L110: "also"?

36 L111: "also"?

37 L116-117: Strange sentence.

37.1 I think it should read “*By* using a model framework …”.
37.2 “robustly”. Really?
37.3 “windows in time” → “time windows”.
37.4 “the skill of index reconstruction” sounds weird.

38 L107-127: This text seems not to belong to the “Methods” section.
39 L120: I don’t think “Alternatively” is the right word here.
40 L122: What do “dust particle records” do here? I’d remove this sentence (and also it seems strange to start it with “Finally” and then you begin the next sentence with “In addition”).
41 L123: “In addition to *this*, …” “this” is very unspecific here. Please rephrase. Also, I’ve notice you use “this” at various places in a similar unspecific way. Please check and clarify where necessary.
42 L125: “… during an entirely different season.” instead of “… during a different season entirely.”?
43 L127: Consistency with what? Why should annual means be more consistent than seasonal analyses? I’m not convinced by this argument.
44 L130: As far as I know, most commonly latitudes between 40°S and 65°S (rather than 60°S) are used for the definition of SAM and only a minority of studies use a different definition. Is there any reason that would speak for 60°S instead of 65°S?
45 L134: I don’t think “established” is suitable here. Maybe “modelled” or something in that direction? Also I’d use plural (“running correlations”).
46 L132 (and whole manuscript): Be concise and consistent throughout the manuscript when using “proxy”, “proxy record”, “proxy data” and “proxy archive”. Here I would maybe use “proxy record”.
47 L132 (and whole manuscript): I just noticed here that you use “the SAM” and three lines later only “SAM”. Please be consistent.
48 L135 (and whole manuscript): Here you write “non-stationary” and on the next line “nonstationary”. Please be consistent.
49 L147: I’d replace “-” with a comma and on the next line there is a “the” missing before “SAM index”.
50 L148-149: I don’t think this sentence is correct. You say the red noise is a combination of random Gaussian noise and lag-1 autocorrelation of a climate variable time-series. But random Gaussian noise is a time-series, whereas the autocorrelation of a time-series is a number. The formula is correct, but the error happened in the attempt to put it into words I think.
51 L152-154 sounds a bit strange / weird formulation. Please rephrase.
52 L165: “metric”? Do you mean “climate variable”. Also “deemed” sounds strange to me.
53 L169: “For this reason” does not sound logic here.
54 L182: Strange sentence. Is there a word missing?
55 L185-188: I’m happy with the choice of CPS as reconstruction method, but I think you can’t say in such a general sense that it is considered to be superior to other methods.
56 L192: Where do you define how your model “nino3.4” index is calculated. I think this should be mentioned somewhere.
57 L194: Where do you introduce the abbreviation “n3.4”?
58 At this point I will stop pointing to typos, inconsistencies and unclear sentences, because, as explained above, I think major revisions of the figures and text is necessary anyway.
59 Maybe a single last comment to Figure 3, its caption and the use of “teleconnection”:
- Do you really to display the legend four times?
- The sentence that starts with “This is therefore …” does not follow from the previous. Also I don’t understand why you say “running” windows.
- Sometimes you use “teleconnection” as substitute for “(running) correlation”. These are generally not necessarily the same. While teleconnections are more “general”, correlations may only capture a specific aspect of a teleconnection.