Dear Eduardo Zorita,

On behalf of the co-authors and myself, we are pleased to submit the revised manuscript. We have addressed all the reviewer's comments – reviewer comments are in bold, while our comments are in normal font. We have included a revised manuscript as well as an improved supplement. We believe that making these revisions has made the manuscript stronger so we would like to thank both reviewers for their constructive comments. The marked up manuscript (using latexdiff) is included after this response.

Regards, Ryan Batehup (on behalf of the authors)

Reviewer Comments #1

Dear Oliver Bothe,

Thank you for your thorough and constructive comments on the manuscript. We have addressed each of the points raised. Your comments are in bold font, while our responses are in normal font. We have attached the revised manuscript (differences highlighted), updated supplementary, and an extra figure in this response.

Major comment

There are two possibilities, either I misunderstand the description of what has been done in the PNEOF1 "experiment", or the "experiment" doesn't represent what it is meant to show.
Maybe what you do is the following: you do the running correlations, you do the EOF, and you then select proxies from the regions deemed non-stationary and also having strong associations with EOF1. If this is the case, the method does what it is meant to do.
However, I understand the description of what you do as the following: you do the running correlations, you do the EOF, and you select proxies from the regions with strong associations with EOF1. Then, you do not sample from the non-stationary regions but possibly from the weakly correlated regions, which would explain the near-total lack of skill for some methods.
Let me rephrase: do you use the EOF to sub-sample the nonstationary regions for covarying regions; or, do you use the EOF to just organise all regions for covariability?

We performed the EOF on the running correlations (our measure of teleconnection strength) to organise all regions for co-variability. Out of extreme values (>abs(0.01)) of EOF1, 16-32% of grid points had a sufficiently high correlation (>abs(0.3)) to be used in the PNEOF1 experiment. We did not restrict the pseudoproxies to be non-stationary as this would have further reduced the available pseudoproxies to create reconstructions.

If my understanding is correct, do you assume, and if so why, that the EOF1 represents nonstationarity?

This is an incorrect assumption that we did make in the submitted manuscript. We thank the reviewer for picking up on this and we have revised the manuscript text to better reflect the analysis that was carried out (see page 19, line 11 of the revised manuscript).

"However, if pseudoproxies are selected from regions that demonstrate co-variability in the running correlations between TS and Niño 3.4 SST anomalies, reconstruction skill is devastated. To this end, an Empirical Orthogonal Function analysis (EOF) was used to 'organise' this co-variability, of which it is expected that non-stationarities are a major part. This is seen in the PNEOF1 experiment shown in Figure 7."

We have computed the percentage of non-stationary pseudoproxies in the EOF to be 9-15%. This suggests that regions displaying coherence in non-stationarities can cause problems with reconstructions even when selecting grid points that are not considered non-stationary. We have now discussed this briefly in the revised manuscript on page 19, line 29.

"The proportion of non-stationary grid points used in the PNEOF1 reconstructions was small, ranging from 9-15%. However, there was still a substantial loss of skill in these reconstructions even though the majority of grid points were classified as stationary by our statistical definition. This implies that a large and coherent change to the teleconnection exists in that region even if it considered mostly statistically stationary, and that was enough to degrade reconstruction skill. Thus, care should be taken to avoid the scenario where all constituent pseudoproxies used in a reconstruction lie in a region where there are large, coherent variations in teleconnections, even if these variations are considered stationary."

The section in the discussion has also been modified (page 27, line 21):
"However, this skill improvement is affected by the degree of non-stationarity and teleconnection co-variability present in the reconstructions, with non-stationary proxy networks NSTAT_ntrop_ts, Fig. 8,red lines) and ``organised'' teleconnection co-variability (PNEOF1, Fig 7 a--d) reducing the degree of improvement in skill with increasing network size. Thus, where increasing network size would usually improve the reconstruction, non-stationarities and spatial coherence in variations in teleconnection strength can substantially temper this improvement. In extreme cases, where proxies are selected from co-varying areas (PNEOF1), Fig. 7), reconstruction skill may show no improvement with larger proxy networks. This further stresses the importance of ensuring that all constituent proxies utilised in a reconstruction are not affected by co-varying teleconnections. This is more likely achieved in spatially diverse, large multi-proxy networks."

Anyway, what does "EOF weighting < 0.1" [Line 18 on page 3871] mean? Do such values occur? (According to the color bar they don't.).

Thank you for pointing out this error. The value '0.1' in the manuscript text should have been '0.01'. This has been corrected in the manuscript.

Additionally, I wonder whether the trend in the PC1 suggests some problems in the control run.

There is little model drift in the GFDL CM2.1, being a 0.1°C rise in Niño 3 region sea surface temperatures (SSTs) per millennium (Wittenberg 2009). Further to this, the PC1 being discussed was computed from the running correlations, not the SST directly, so it does not necessarily indicate model drift as we can assume any drift is coherent across all grid points, which would not manifest in a change in correlations. This model is shown to "exhibit strong interdecadal and intercentennial modulation of its ENSO behavior" (Wittenberg 2009), and we believe what is seen in PC1 is simply due to some of this intercentennial ENSO modulation.

Minor comments:

1**. Could you please as soon as possible, i.e. before your final response to the reviewers, provide a version of Figure 10 including the MRV?**

Figure 10 showed a few hand-picked reconstructions to show that there are some instances where variability (standard deviation) between proxy time series appears to be related to variance losses in the final reconstruction. We were unable to produce a new figure that accurately summarised the ideas of Figure 10 and removed the reviewer's confusion. Thus, we have decided to drop this figure and its corresponding paragraph in the manuscript. This manuscript change does not influence the final results of the manuscript.

2. I think the title should mention that you reconstruct of ENSO-variance. The introduction also should clearly state it. Similarly, on some/most instances where you write "reconstructions of ENSO" it would be more appropriate to write "reconstructions of ENSO variance". Alternatively you may state early on that "reconstruction of ENSO" implies "reconstructions of ENSO variance".

The following changes has been made:

The title has been changed to "The influence of non-stationary teleconnections on paleoclimate reconstructions of ENSO variance using a pseudoproxy framework".

The first sentence of the abstract's second paragraph has been changed to "This study examines the implications of non-stationary teleconnections on modern multi-proxy reconstructions of ENSO variance".

The first sentence of the abstract's third paragraph has been changed to "We find that non-stationarities can act to degrade the skill of ENSO variance reconstructions. However, when global, randomly-spaced networks (assuming a minimum of approximately 20 proxies) were employed, the resulting pseudoproxy ENSO reconstructions were not sensitive to non-stationary teleconnections".

In various places: "ENSO variability" has been changed to "ENSO variance".

Similarly at Page 3868 Line 20ff: The description of Figure 4 is, as far as I understand it, incorrect. Please be clear that it is the correlation between the running variance series and not between the Nino3.4-indices.

The sentence has been changed to: "the correlation between the pseudoproxy reconstruction of the Nino 3.4 running variance and the model Nino 3.4 running variance" (on page 16, line 22).

3. Methods:
a. Is the calibration window length implicitly meant to also be the length of the running variance windows? I ask, because you never explicitly mention the window-length for the running variances. It can't be the calibration window length, because you also use the full 499 years for calibration. So, what is the window length for the running variances?

The running variance window length is 30 years for all experiments. We have added the sentence: "All running variances were calculated using 30 year windows." to the revised manuscript to clarify (see page 14, line 4 of the revised manuscript).

b. Please clarify what calibration means in your setup.

Calibration is similar to the normal calibration between the paleoproxy time series and the instrumental record. In our experiment, calibration establishes relationships between the TS grid points and the model Nino3.4 index during a certain time period, which is the calibration window. We have inserted the sentence "The calibration window is the time period where relationships between the TS grid points and the model Nino3.4 index are established" on page 9, line 17.

c. You mainly consider skill in terms of correlation. Is correlation really the best skill measure in this case? Why?

We use correlation mainly for simplicity. We do also look at RMSE, but it is not within the scope of the study to examine the best skill measures for paleoclimate reconstructions. If it is an either correlation or RMSE choice, we would argue the correlation is the most important metric. For instance, what good is a small error if you have no idea whether the variance is increasing or decreasing? It is much more useful to know the relative changes, and to understand how what we are seeing now relates to what has happened in the past.

d. On page 3860 line 10 you describe the timeseries as "June-July"-averages. I assume you mean thirteen month averages. Please clarify.

This meant to mean 12 month averages, from July to June of the following year. This typo has been corrected.

e. You mention model-drift: Is it a problem in your simulation?

No, we do not think that model drift is an issue in this simulation. For instance, the study of Wittenberg (2009) who produced this data reports a drift of only 0.1°C per millenium in the GFDL CM2.1 eastern equatorial Pacific SSTs (Wittenberg 2009).

f. I wonder whether it would be better to convert all correlations to Fisher-Z-scores. I do not propose to do it, I only wonder whether it might clarify the Figures.

It would help to distinguish differences at higher correlations, but we feel leaving it as correlations is simpler and easier for the reader to understand. As such, we have left it as correlations.

4. There are references which should not be omitted. A paper on reconstructing ENSO-variance has to mention the work by Russon and colleagues (Russon et al., 2014, 2015). Furthermore, the general introduction of ENSO-reconstructions appears to ignore the works of Watanabe et al. (2012), Cobb et al. (2013), and Li et al. (2013).
As a less important side-note I want to mention that there have been other studies dealing with non-stationarity of climate-modes in recent years.

A few sentences have been modified on page 4, line 15: "Tropical corals are the dominant proxy type in this region, and are known to provide very skilful reconstructions of the surrounding SSTs and ENSO. However, the addition of non-climatic noise to these proxies also complicates the estimation of the significance of changing in past ENSO variability (Russon et al. 2014; 2015), as does their limited life span (i.e., records are on average about 50 yrs in length, with the longest records less than two centuries) (Cobb et al., 2013; Neukom and Gergis, 2012)."

- Russon et al 2014 and 2015 has been inserted into page 4, line 19 (as above).
- Cobb 2013 has been inserted into page 4, line 22 (as above).
- Li et al. 2013 has been inserted into page 4, line 22.
- Watanabe2012 has been inserted into page 4, line 7.

5. On the better performance of MRV.
a. I think this would be an interesting test for what happens when noise is introduced in the proxies.

It would be interesting to explore this, but it is beyond the scope of the current paper. However, we do intend to examine this in future research.

b. You stress the overall better performance of MRV. One may argue that this is unsurprising as it uses the variance from the beginning, in this sense the comparison may be called biased.

We do not believe that this is a bias as the difference between the MRV and RVM method is simply due to the order of operations, which is not obvious.

However, on the other hand, I am not sure it is true. In a real world scenario: what is the more important skill-metric, the RMSE, where MRV performs consistently worse, or the correlation? Is there possibly another better suited skill-metric? You mention the potential need for re-scaling but don't do it. Why not, I would think implementing an MRVPS (MRV plus scaling) should be easy enough. There may be reasons, but as you so far don't show the MRV series (e.g., in Figure 10), the reader is unable to assess this.

The question of which is the "correct" metric with which one should assess reconstruction skill has been widely debated in the literature for a long time. We do not intend to assess different metrics in detail here, as it is not the main point of the paper. But in response to the first question raised here, if it is an either-or choice, we would argue the correlation is the most important metric. For instance, what good is a small error if you have no idea whether the variance is increasing or decreasing? It is much more useful to know the relative changes, and to understand how what we are seeing now relates to what has happened in the past (from comment 3c).

Before moving onto the second component of this question, we would like to highlight that each of the reconstruction methods utilised, suffers from variance reduction. A feature that is clear in Figure 10 and Supplement Figure S6 of the revised manuscript, and is discussed on page 22, line 23.

"It is well known that all reconstruction methods result in a loss in ENSO variance, and this is clearly shown in Figure 10. In Fig.10a-d, we can see that, all reconstructions underestimate the model Nino 3.4 running variance (black line). However, this figure also shows that this variance loss is exaggerated with the MRV method (panels c, g), and this is also seen in Supplementary Fig. S6, particularly at the larger network sizes."

We expect, and will show here (and in the revised manuscript), that scaling the resulting running variance time series improves the RMSE of the MRV, but also note that it cannot be done well when limited to 31-yrs of data (the 31-yr calibration window). Thus, we have performed scaling on the 61 and 91-yr calibration window experiments. We have scaled using the average (calculated over the 1000 reconstructions) regression of the MRV and Nino3.4 RV within the calibration window to produce a scaled MRV – 'SMRV'. The proportions of the SMRV with lower RMSE than MRV is up to 70% and is indicated in the attached Figure E1, with the proportion getting larger as the number of proxies in the reconstructions increase. The median RMSE of each reconstruction method is also indicated on the plots on Figure S7.

We have also discussed this scaled MRV in a paragraph on page 23, line 25 and added a Figure S7 to the supplementary (while incorporating Fig. S3 into a new Fig. S2).

"In order to compensate for the variance loss of each reconstruction (Figure 10a-d), we rescale each method's resulting running variance time series (Figure 10e-h). Rescaling the running variance time series was carried out using the average (calculated over 1000 reconstructions) regression between the reconstructions and the modelled Nino 3.4 running variance within the calibration window. When the MRV (panel 10.c) is scaled to form the SMRV (panel 10.g), there is a jump in reconstruction variance (grey shading), such that the modelled Nino 3.4 index running variance is now encompassed by the grey shading. Using this simple scaling technique, we see a large reduction in the RMSE (see Figure S7) - up to a 0.1 reduction in the median (Figure S, panels c, g, black line) and no changes in the correlation (not shown). In fact, it is noteworthy that on average the scaled MRV has the smallest RMSE (significant to 99% level via a two sample t test) of all reconstruction methods."

Note that examples of the MRV reconstructed running variance are now provided in the new Figure 10 mentioned above, showing the range of the reconstructions compared to the Nino3.4 RV series they are attempting to reconstruct.

c. Similarly, page 3875 line 13ff: Isn't the damping an expression of especially large variance loss for MRV, at least if the term is used as commonly employed for reconstructions?

Yes, as described in our response to your point above, the MRV displays an exaggerated variance loss as the term is currently used. This is now discussed in the revised manuscript on page 23, line 4.

d. Page 3877 line 21: You write MRV excelled, but MRV also showed large RMSE. I think that should be mentioned. Can you estimate how noise/uncertainties in the proxies would affect this feature.

In regards to the latter point: as stated above, to examine this would require looking at the effects of noise in the pseudoproxies which is beyond the scope of this study. However, we do highlight the need for this to be examined in future research: "The compounding effects of noise and non-stationarities on the reconstruction method and hence, a reconstruction, should be the focus of future research efforts in this area" on page 26, line 18.

In regards to the former point: We have made some changes to better reflect the large RMSE: "However, the unscaled MRV method showed poor RMSE performance, meaning that it can only be used to provide useful information on the relative changes in ENSO variance." has been inserted into page 27, line 3.

The first line of the second paragraph of page 22 (line 17) has been changed to: "It is worth noting that although the MRV method shows the most consistently high correlations to ENSO and appears to be the least sensitive to calibration window position (smallest percentile ranges, Supplement Fig. S5), it has the highest RMSE (root-mean-square error)". A scaled version of the MRV, the 'SMRV' has been created - see previous comments.

e. Page 3879 line 16/17: MRV is the most robust in your perfect-proxy setting.

Again, we now discuss this possible limitation of our results. See page 27, line 3, where we added the sentence: "However, the unscaled MRV method showed poor RMSE performance, meaning that it can only be used to provide useful information on the relative changes in ENSO variance."

6. With respect to the discussion of CPS on Page 3873 Line 15ff. Why do you single out CPS here? It is not really worse than RMV or EOF.

We were actually highlighting the CPS as the next best method, due to this confusion we have now reworded this sentence from “It is noted that the CPS_RV method performs well, although mainly with longer calibration windows and for the random selection experiments” to “The CPS_RV method performs almost as well as the MRV, although mainly with longer calibration windows and for the random selection experiments” on page 21, line 26.

7. On the discussion of the tropical supposedly non-stationary grid-points on page 3865: Don’t these mainly represent the wide range of internally varying differences in the evolution of ENSO-events which are potentially not captured by a simple stochastic process? I am not so much thinking of CP vs. EP ENSO, but different evolutions of one of these flavors. I think this could be discussed more extensively. You make a similar point on page 3870, line 15ff. So, it’s not only about different flavors but about the large variability in how events evolve. The statistical process captures not necessarily all dynamical variability.

We have discussed this more extensively on page 13, line 16:

"Of further note is a large non-stationary area in the equatorial Pacific; given this is the area surrounding our ENSO index it is debatable whether this should be considered as a non-stationarity. Rather, we expect the changing relationship in this surrounding region to be the result of complexities of ENSO that may not captured by the simple stochastic model of stationarity. For instance, ENSO displays: i) significant non-linearities in its magnitude (An and Jin, 2004) and duration (Okumura and Deser 2010); ii) differences in the evolution of events with all La Nina's and most small to moderate El Nino's having SSTAs that propagate from east to west, while the SSTA of large El Nino events propagate from west to east (Santoso et al. 2013); and iii) changes in its spatial structure (CP-EP type events) which may be considered different flavours of events rather than non-stationarity teleconnections of the event (Gallant et al., 2013; Sterl et al., 2007)."

8. On Figure 10: The regressions and the correlations show rather weak relations between the two series. Anyway, why should the standard deviation of the running correlations of the proxies with the target and the variance reconstruction be related in any way. Put differently: what do we learn from the correlation of the variance of a measure with the square-root of the variance of running correlations of a measure with another measure?

We were examining the possibility of the variability between the proxies being related to abrupt changes in the reconstructed variance. As discussed in our response to this reviewer’s first minor comment, we were unable to produce a new figure that accurately summarised the ideas of Figure 10 and removed the reviewer’s confusion. Thus, we have decided to drop this figure and its corresponding paragraph in the manuscript. This manuscript change does not influence the final results of the manuscript.

The number of effective degrees of freedom of the time series appears to be small, are the correlations even significant?

Figure 10 has been removed from the revised manuscript - see minor comment 1 and the above comment.

10. In your plots you give explained variances for running variances. How much variance is captured by the stationary sample-variance-distribution?

We have looked at the distributions of the reconstructed running variances, and they closely fit with a normal distribution for all reconstructions, which suggests discussing the explained variance is valid. However, we are not sure if we understand the question correctly. We also would like to refer to Fig. 5 for a comparison of stationary and non-stationary reconstruction skill.

11. The discussion in the beginning of the last paragraph on page 3872 is not really relevant, is it? The case of a 500 year calibration window is not realistic, so the discussion should focus on the skill differences between 91yr and 31yr. Related: The second part of the paragraph (on page 3873) appears to be partially redundant.

The reconstruction skill is affected by both non-stationarities and reduced window size. Thus we believe it is fair to look at the impact of reduced window size alone, before making a judgement on the impact of non-stationarities. We have also removed the redundancies.

12. Page 3876 line 2ff: Please do not just point to panels but give some more information on what we see.

The following changes have been made to the surrounding text:

Page 24, line 24: "absence of a large spatially coherent region of correlations in the tropical Pacific Ocean (see Supplement Fig. S1e)" has been changed to: "absence of a large spatially coherent region of correlations in the tropical Pacific Ocean (compare tropical areas in Supplement Fig. S1b and Supplement Fig. S1e)"

Page 25, line 6: "The RVM method appears to perform better with precipitation than temperature in panels d, and h, with not much difference in panel l, which is consistent with the findings of McGregor et al. (2013)" has been changed to "The RVM method appears to perform slightly better with precipitation than temperature (Supplementary Fig. S2, mainly at longer calibration windows), which is consistent with the findings of McGregor et al. (2013)."

13. It may help the reader if you extend a bit in line 22ff of page 3876 on what Wittenberg (2009) showed.

The sentence on line page 25, line 27 has been changed to: "Wittenberg 2009 discussed that such changes to ENSO behaviour could conceivably alter the teleconnections between ENSO and local climate and that these changes may not be represented in the historical record."

14. More a technical comment, but I put it here as well to emphasize it: I think Figure 8 is rather unclear, since I am not really able to distinguish the different hatchings.

Figure 8 has been simplified and clarified. The hatching has been removed and the remaining lines of proportion have been compressed into four panels. See the Figure 8 in the updated manuscript.

Technical Comments

General: I am surprised by the comma placement in the manuscript. However, as a non-native English-speaker, I do not annotate it.

The comma positions have been corrected where necessary.

My subjective impression is that the abstract could be shortened and could formulate the relevant points more concisely.

The abstract has been shortened and made more concise.

Page 3854 Lines 17-20: Please rephrase the sentence. (I do not really see how the "to which"-part relates to the previous sentence-structure.)

This has been changed to "Reconstructions of the variance in the Niño 3.4 index representing ENSO variability, were generated using four different methods. Surface temperature data from the GFDL CM2.1 were used as pseudoproxies for these reconstruction methods" on page 2, line 17.

Page 3854 Line 24: What do you mean by "uniformly-spaced"?

This has been changed to "randomly-spaced" on page 2, line 27.

Page 3859 Lines 21-24: I think this sentence could be clarified "It has also been shown that the model teleconnections, represented by correlations in 31yr windows between grid points and the Nino 3.4 index generated from the model, do exhibit variability between periods and compared to correlations calculated over the entire period (Fig.~\ref{fig1}a, \citealp{Wittenberg2012})."

This sentence has been changed to "It has also been shown that the model teleconnections, represented by correlations in 31 year windows between grid points and the Nino 3.4 Index generated from the model, do change over time, and differ compared to correlations calculated over the entire period (Fig. 1a, Wittenberg, 2012)" on page 7, line 29.

Page 3860 Lines 21-22: "is used" : : : "are selected". Please clarify is/are.

This has been corrected to 'is' on page 9, line 3.

Page 3862 Line 3: "Fig. 1b" -> "(Fig. 1b)"

The correction has been made on page 10, line 10.

Page 3864 Line 22: you write of the "possible" range of running correlations. Is "possible" the correct word here

We believe it is: "A 95% confidence interval was generated at each grid point from the stochastic simulations and was used to represent the range of running correlations possible, assuming a teleconnection was stationary" on page 12, line 24.

Page 3865 Line 22: Do you examine the "likely" - as you write - or the "potential" effects of non-stationarities?
This has been corrected to 'potential effects' on page 14, line 2.

Page 3866 Line 22: "(2005); Hegerl" -> "(2005) and Hegerl"

This has not been changed as we believe that there are too many references to place an 'and' there. It is currently: "(Esper et al., 2005; Hegerl et al., 2007; Mann et al., 2007, and references therein)."

Page 3867 Line 16: I don't think you write in Sect. 2 how you calculate the running variance for the ENSO index (see above, what is the window-length of the running variance).

We have added a line at the beginning of 3.3 Reconstruction Methods: "This study examines the potential effects of non-stationarities on multi-proxy reconstructions of the running variance of the Nino 3.4 index (representing the variability of ENSO) using pseudoproxy data. All running variances were calculated using 30 year windows" on page 14, line 4.

Page 3867 Line 19: “dataset is available with larger proxy networks”. My impression is, that there may be an “and” missing between available and with.

We have corrected this to: “when the entire dataset is available (and with larger proxy networks)” on page 15, line 24.

Page 3869 Line 5: “The skill metrics”, I think you mean “the proportion of skill metrics”?

The blue and orange lines in Figure 4 are used to evaluate the skill of the methods. Although they are indicating the proportion of skilful reconstructions (skilful meaning explaining greater than 50% of variance), this proportion itself can be used as a skill metric for comparative purposes. See page 17, line 8 in the revised manuscript.

Page 3870 Line 1: “In all”. Should that be “For all”?

Yes, that has been corrected on page 18, line 3.

Page 3872 Line 15ff: I think the second part of this sentence is incomplete.

This sentence has been removed.

Page 3872 Line 18: I don’t mind “Fairly good chance” but I can imagine colleagues who are rather annoyed by such a phrase.

We have corrected this to: “using a minimum of 20 proxies gives a reasonable chance” on page 21, line 1.

Page 3872 Line 26: Is “would be” correct? Shouldn’t it read “is” which may be qualified by “likely” or “potentially”.

We have corrected this to “This decrease of skill is potentially due to some information loss in the relative datasets, and not necessarily due to non-stationarities” on page 21, line 8.

Page 3873 Line 7: I think the “However” is wrong.

This sentence has been removed.

Page 3873 Line 15: “It is noted”? Why passive construction? -> You note.

This sentence has been changed (page 21, line 26), see minor comment 6.

Page 3874 Line 3: Not the red line outperforms the other lines, but using 91 year calibration windows performs better than shorter windows.

Corrected to: “and displaying some sensitivity to calibration window length (91yr windows perform better than shorter windows)” on page 22, line 14.

Page 3874 Line 16ff: Is the sentence correct? Do you generally plot the “variance taken : : : of the correlations”?

Yes, this is the variance of the reconstructions. As the reconstructions are running variances themselves, this is the variance of the running variance. However, this sentence has been removed in the revised manuscript.

Page 3875 Line 18: Is "on this paper" correct?

This has been corrected to "Although not the focus of this paper, precipitation was also examined for all experiments" on page 24, line 18.

Page 3877 line 15: "highlight a case for considering" or just "highlight"?

We believe the existing sentence is more suitable: "the results presented here highlight a case for considering the influence of non-stationarities on real-world reconstructions and their underlying methods" on page 26, line 23.

Page 3878 line 21ff: I think the second part of this sentence is incomplete.

"Given the skilful reconstructions in ENSO variance that can be produced by neglecting pseudoproxies from the centre of action, as shown here, the utilisation of data solely from the eastern equatorial Pacific appears unnecessary" has been corrected to:

"Given the skilful reconstructions in ENSO variance that can be produced by neglecting pseudoproxies from the centre of action as shown here, the utilisation of data solely from the eastern equatorial Pacific appears unnecessary" on page 28, line 6.

Page 3879 line 17: "many various"?

We have corrected this to: "and there are many reconstruction methods in the literature" on page 29, line 7.

Page 3879/3880: My impression is that the second part of the conclusions is more or less redundant and repeats what the first part already said.

The redundancies have been removed (see page 29, line 12).

Page 3880 line 5ff: I think this perspective is not really necessary, but that's just personal taste.

This line has been removed.

Figure 1: Please rephrase "is the correlation between of the entire 499 years of TS at each grid point and the model calculated Nino 3.4 index correlation coefficients"

This has been corrected to: "The shading is the correlation between of the entire 499 years of TS at each grid point and the model calculated Nino 3.4 index, both calculated from the GFDL CM2.1 data" (page 36).

Figure 2: Please provide labels for the color bars of panels b,d,f. verses -> versus

A new figure has been prepared – see attached updated manuscript (page 37).

Figure 3: You write "the pseudo-reconstructions running variance". Wouldn't it be more appropriate to write the "pseudo-reconstructions of running variance"?

Yes, it would be. I have corrected this to "between the pseudo-reconstructions of running variance and ENSO running variance" on page 38.

Figure 4 (and other Figure captions): You write "reconstruction's running variance" which is appropriate for CPS and EPC and to some extent for RMV, but shouldn't it be "reconstructed running variance".

Yes, thank you for pointing that out. This has been corrected in Figures 4, 5, 8 and 9.

"explaining greater than 50% of explained variance" -> "explaining greater than 50% of variance"?

This has been corrected to "explaining greater than 50% of variance".

Figure 6: Please rephrase "and this determines what values of proportion can be taken as larger groups have a wider range of possible non-stationarity proportions than smaller groups".

This has been changed to "determines what values of proportion can be taken, hence larger groups have a wider range of possible non-stationarity proportions than smaller groups" on page 41.

Figure 8: The hatching is, from my point of view, nearly unidentifiable.

The hatching has been removed in the new Figure 8.

Supplement: Please provide a clear copyright statement.

There will be a copyright statement on the front page of the supplementary material when it is processed by the editor.

Figure S1: Could you reshape the aspect ratio for final publication?

The aspect ratio has been fixed.

Figure S4: Caption, last line: "with with"

The repeated word was removed.

Reviewer Comments #2

Dear Anonymous reviewer,

Thank you for your thorough and constructive comments on the manuscript. We have addressed any issues that have been found. Your comments are in bold, while our responses are in normal font. We have also attached the revised manuscript (differences highlighted), updated supplementary, and an extra figure in this response.

Comments:

Introduction: The last paragraph leaves the reader thinking that teleconnection nonstationarity does not impact reconstruction skill. This is not the case and I would start this paragraph by noting that non-stationarity does degrade reconstruction skill. You can then follow that by

noting that the impact of non-stationarity can be minimized by employing a large global network.

We have left the discussion of results out of the introduction as we decided to stick to the traditional report format. As our paper explores the impact of non-stationarities on the reconstruction skill, we have left these questions open ended in the introduction. However, we believe that the reviewer may be referring to the last paragraph of the abstract, in which case we have now reworded the opening sentence:

"We find that non-stationarities can act to degrade the skill of ENSO variance reconstructions. However, when global, randomly-spaced networks (assuming a minimum of approximately 20 proxies) were employed, the resulting pseudoproxy ENSO reconstructions were not sensitive to non-stationary teleconnections".

Page 3854, Line 15: add variance after ENSO. More generally, this is a problem throughout the manuscript. Make sure to be clear that these are reconstructions of ENSO variance. I will note places where this should be clarified but have likely missed some.

The fact that we are reconstructing ENSO variance has now been clarified throughout the revised manuscript.

Page 3854, Line 27: This sentence seems tangential to the overall results and perhaps overly specific.

This sentence is intended to describe our result that the MRV reconstruction method appears to perform better than other methods. We have reworded this sentence to: "Different reconstruction methods exhibited varying sensitivities to non-stationary pseudoproxies, which affected the robustness of the resulting reconstructions" on page 3, line 4.

Page 3855, Line 11: Suggest removing: "Thus, these proxies are the essential tool for creating paleoclimate reconstructions."

This sentence has been removed.

Page 3857, Line 28: Suggest removing: "This places increasing stress on the assumption that teleconnections are stationarity. Further to this, it."

We believe this sentence is important to tie in the multiple results from different papers and show their relevance to our manuscript. This sentence has been tweaked to "This places increasing stress on the assumption that teleconnections are stationary. This raises the question as to whether non-stationarities have an appreciable influence on the robustness of past paleoclimate reconstructions" on page 6, line 9.

Page 3858, Line 4: I would replace variability with variance to be absolutely clear that that is what is being reconstructed.

This has been done.

Section 2: You are only using a single model and it will be important to note that PPE results have been shown to be model dependent, at least in the case of CFRs (e.g. Smerdon et al. GRL 2011 and Smerdon et al. Clim. Dyn. 2015). This is particularly important given that not all models have non-stationary teleconnections to the tropical Pacific (e.g. Coats et al. GRL 2013). Using a different model may provide different results, so making absolutely clear that the results are specific to the characteristics of this GFDL model and not necessarily applicable to the real world will be important (after all, a model with either stationary teleconnections or much more non-stationary teleconnections is arguably an equally plausible representation of the real world).

The reviewer is correct as the global number of non-stationary grid points may vary depending on the climate model. However, we believe that this is unlikely to affect the differences between reconstruction methods as these have been examined using only source pseudoproxies that can be considered non-stationary (see Fig. 5 of the revised and original manuscript). In order to address the reviewers concern, we do now discuss this potential caveat on in the conclusions on page 28, line 17: "These results make the implicit assumption that the modelled co-variability of the non-stationarities and relative proportions of non-stationary areas to stationary areas are realistic, which has not been explicitly tested here". The sensitivity of multi-proxy climate index type reconstructions to the climate model will be examined in the future though.

We have modified a sentence in the discussion: "We note that although we use the same model as in the Wittenberg (2009) study, the results are unlikely to be a product of the model configuration given that Gallant et al. (2013) identified nonstationarities in three different GCMs" has been changed to:

"Gallant et al. (2013) identified non-stationarities in three different GCMs. It is noted that while numerous models display non-stationarities, their regional existence may vary depending on the model used (Coats 2013). We do not expect our evaluation of various different reconstruction methods performance in the presence of non-stationarities to be affected by model configuration, however we intend to examine this in future research" on page 26, line 5.

Page 3860, Line 10: Does June-July imply a two month average? Based on the rest of the sentence I assume you mean a 13 month average. Please clarify.

This was a typo, and has been corrected to "July – June" 12 month averages on page 8, line 17.

Page 3860, Line 11: The sentence on computational cost seems unnecessary.

This sentence has been removed.

Page 3861, Line 4: The reference to Lee et al. (2008) seems out of place because adding non-climatic noise at different levels is a relatively standard choice in PPEs. Perhaps restructuring to put the reference at the end of the sentence with a parenthetical note that Lee et al. is an example of this. The reference is also not listed in the references.

The sentence on page 9, line 9 has been changed to: “The pseudoproxies are not degraded by adding noise (e.g. Lee et al. 2008), as the effects of noise on the reconstructions are beyond the scope of the study”.

The reference has been added to the reference section.

Page 3861, Line 11: Suggest removing “to some extent, making them at least partly relevant for reconstructing the ENSO signal.”

As suggested, the sentence “This threshold is an arbitrary criterion that is simply there to ensure the pseudoproxies represent ENSO to some extent, making them at least partly relevant for reconstructing the ENSO signal” has been changed to “This threshold is an arbitrary criterion that is simply there to ensure the pseudoproxies at least partially represent ENSO” on page 9, line 14.

Page 2861, Second paragraph: Are 1000 random networks of each size from three to 70 used?

Yes, the reviewer is correct. We have now modified the sentence on page 9, line 23 to read: “To produce reconstructions of the model Nino 3.4 index variance, 1000 random networks were selected per network size, calibration window length, and calibration window position. “

Page 3861, Last paragraph: It is important to note that a real reconstruction will only be able to calibrate on the observational record. The first sentence, however, seems to distract from this important point - perhaps try rewording.

“The correlation at each grid point over the whole time period (499 years) and ENSO is assumed to represent the true teleconnection strength, as its use for calibrating the proxies should result in more accurate reconstructions” has been reworded to “The correlation between ENSO and each grid point time series (i.e. Nino3.4 & TS) over the whole time period is assumed to represent the true teleconnection strength” on page 10, line 1.

Page 3863, Line 5: The last sentence is unclear and doesn’t seem necessary.

The sentence “However, our experiments showed that this assumption also produces larger errors in the reconstruction (not shown)” has been removed.

Page 3863, Line 8: But you do show results from these experiments in the other figures. Perhaps remove this sentence.

We thank the reviewer for highlighting this discrepancy. The sentence “For the remainder of the paper, we show the second version of the experiments only, as it represents the most realistic case.” has been changed to “As the second case is the most realistic case, we mainly focus on the second version of experiments for the remainder of the paper” on page 11, line 12.

Page 3865, Line 8: Do you mean segments or windows and not years? A year can’t be non-stationary but the 31-year window, for instance, can. Or are you counting up the number of years within all the non-stationary windows? That would be much less intuitive. If it is the former I would suggest changing years to windows or segments and doing the same in the corresponding figure and caption.

Yes, we do mean windows in this instance as each year we were discussing implicitly includes the 15-year period either side of the year in question. Thus we have simply changed 'year' to 'window' (see page 13, line 10). The caption of figure 2 has also been updated to be consistent with the manuscript text.

Page 3868, Line 21: I think that you mean the running variance of the Niño3.4 index.

This has been corrected in the manuscript.

Page 3869, Line 13: Perhaps provide a value in parentheses here (after larger network sizes).

The mean percentage of skilful reconstructions has been calculated to be 68%. Using larger network size only (20 to 70), we get a value of 77% of all non-tropical reconstructions being skilful (as according to the manuscripts definition of skilful – explaining more than half the variance of Nino3.4). This value of 77% has been inserted into the manuscript on page 17, line 15 as was the definition of larger network sizes.

Page 3869, Line 17: Again change variability to variance.

This has been corrected in the manuscript.

Page 3870, Line 11-14: How are the psuedoproxies non-stationary if they display little variability in correlation, that is not intuitive.

This was an interesting find amongst the experiments. Since the non-stationary definition doesn't require there to be minimum magnitude change in teleconnection strength, the pseudoproxy time series just needs to deviate from the generated stochastic range earlier defined. This may be a flaw of the non-stationary definition, but rather we believe it is picking up the non-linearities of ENSO events in that tropical Pacific region. This is now discussed in more detail in the manuscript on line 16 of page 13 in response to one of the other reviewer comments.

Page 3870, Line 19: I understood what you were saying after reading further but was confused by this sentence initially. Perhaps try to make the statement more clear.

The sentence "In regards to why non-stationarities do not seem to impact the high skill of random pseudoproxy selection of Sect. 4.1, we find that the likelihood of selecting non-stationarities is relatively low." has been changed to:

"The fact that we see a minimal effect of non-stationarities in the randomly selected pseudoproxy experiments may be because the likelihood of selecting non-stationarities is relatively low" on page 18, line 19.

Page 3871, Second paragraph: I found this confusing. I think what you are saying is that if the network consists of a large proportion of grid-points chosen due to spurious correlations in that calibration window, the reconstruction skill is very low. The statement: "non-stationarities at the same time" might be part of the problem. A grid point is either non-stationary or not, but if

it is non-stationary and weakly correlated it won't always be eligible to be picked and that appears to be what you are getting after.

If a particular window's correlation lies outside the 95% confidence interval generated by the stochastic simulations, then we deem it as a non-stationarity. EOF analysis was employed to determine covariation in non-stationarities and we showed that this covariation negatively effects reconstruction skill (non-stationary grid points made up 9-15% of the PNEOF1 experiment). This paragraph has been changed substantially to address an issue of co-varying teleconnections in response to reviewer 1 (Oliver Boothe) Major comment on page 19, line 11.

Page 3873, Line 10: The result that increasing the length of the calibration window is less important for reconstruction skill as compared to the choice of method or the amount of non-stationarity is important and gets lost a bit in the manuscript.

Thank you, we have now tried to clarify this finding by adding a sentence in the discussion "The non-stationarities and reconstruction method usually had a greater influence on reconstruction skill than the calibration window length" on page 26, line 26.

Page 3874, Line 7: "correlations to ENSO variance." It might be worth changing everything to ENSO variance or everything to Niño3.4 variance throughout the manuscript for clarity.

This has been changed throughout the revised manuscript.

Page 3874, Second paragraph: The discussion of RMSE versus correlation for MRV seems unnecessarily drawn out (and slightly confusing). The takeway appears to be that the scaling of the variance for he MRV method is too low but the timing of variance changes are correct.

This paragraph has been changed in response to reviewer 1's (Oliver Boothe) major comment to include information about variance losses in the reconstruction method on page 22, line 23, along with a discussion of the effect of rescaling the pseudo proxy reconstructions. We have also tried to simplify the discussion to make the take home message clearer (see page 27 line 12)

Page 3877, Line 2: Coats et al. GRL 2013 showed model dependence of non-stationarity to ENSO so this statement isn't strictly correct.

That small section on page 26, line 5 has been changed to:

"Gallant et al. (2013) identified non-stationarities in three different GCMs. It is noted that while numerous models display non-stationarities, their regional existence may vary depending on the model used (Coats 2013). We do not expect our evaluation of various different reconstruction methods performance in the presence of non-stationarities to be affected by model configuration, however we intend to examine this in future research"

Page 3877, Line 3: Suggest removing virtual.

"Virtual" has been removed on page 26, line 11.

Page 3877, Line 26: To this reader it was not obvious why the filtering produced these different interpretations.

We were trying to reconcile differences with previous studies here. We believe that the difference is due to the fact that the unfiltered time series contains decadal signals that are less likely to suffer from the signal cancelation of higher frequency variability, hence they act to enhance the skill of the RVM reconstruction. However, as this has not been thoroughly tested and added little value to the current manuscript, we choose not to elaborate on this further.

Page 3878, Line 1: The point here is not that multi-proxy networks will produce more informative reconstructions, it is that larger and more global networks will. Maybe flip the sentence structure so that in the back part you can explain that multi-proxy networks tend to be larger and more global.

The sentence "For reconstructions of large-scale phenomena like ENSO, networks will produce more informative reconstructions because the larger networks contain more information, including spatial information, compared to single site (Mann, 2002; Lee et al., 2008; von Storch et al., 2009; McGregor et al., 2013)" has been changed to:

"For reconstructions of large-scale phenomena like ENSO, larger more globally diverse networks will produce more informative reconstructions compared to those derived from smaller regions or single sites (Mann, 2002; Lee et al., 2008; von Storch et al., 2009; McGregor et al., 2013). The experiments conducted here support this hypothesis, as the proportions of skilful reconstructions increase as the number of source proxies increase for almost all reconstruction methods and calibration window lengths (Figs. 8 and 5)" on page 27, line 12.

Figure 1, Caption: On Line 7 remove correlation coefficients.

The typo has been removed.

Figure 2: As noted above, years or segments. Segments would be much more intuitive. Colorbar has no label.

This has been fixed in the manuscript.

Figure 10, Caption: put tilda in Nino3.4.

This has been done in the manuscript. However, note that Figure 10 has been removed and replaced with another figure.

Figure 10: I find this figure hard to interpret in the context of the results. The blue line is the standard deviation of the correlations for a 30 year window, where the correlations are between a pseudoproxy and the actual Niño3.4 index. I don't see the relevance but perhaps I am misunderstanding what is shown. The red circles are just a very small subset of the plot, is this relationship consistent? In any case, this figure and the discussion would benefit from further clarification.

This figure has been removed.

Manuscript prepared for Clim. Past Discuss.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 15 November 2015

The influence of non-stationary ~~ENSO~~ teleconnections on paleoclimate reconstructions of ~~paleoclimate~~ ENSO variance using a pseudoproxy framework

R. Batehup[1,2], S. McGregor[1,2], and A. J. E. Gallant[3,2]

[1]Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia
[2]ARC Centre of Excellence for Climate System Science (ARCCSS), Australian Research Council, Australia
[3]School Earth, Atmosphere and Environment, Monash University, Victoria, Australia

Correspondence to: S. McGregor (shayne.mcgregor@unsw.edu.au)

Abstract

Reconstructions of the El Niño-Southern Oscillation (ENSO) ideally require high-quality, annually-resolved and long-running paleoclimate proxy records in the eastern tropical Pacific Ocean, located in ENSO's centre-of-action. However, to date, the paleoclimate records that have been extracted in the region are short or temporally and spatially sporadic, limiting the information that can be provided by these reconstructions. Consequently, most ENSO reconstructions exploit the downstream influences of ENSO on remote locations, known as teleconnections, where longer records from paleoclimate proxies exist. However, using teleconnections to reconstruct ENSO relies on the assumption that the relationship between ENSO and the remote location is stationary in time. Increasing evidence from observations and climate models suggests that some teleconnections are, in fact, non-stationary, potentially threatening the validity of those paleoclimate reconstructions that exploit teleconnections.

This study examines the implications of non-stationary teleconnections on modern multi-proxy reconstructions of ENSO variance. The sensitivity of the reconstructions to non-stationary teleconnections were tested using a suite of idealized pseudoproxy experiments that employed output from a fully coupled global climate model. Reconstructions of the variance in the Niño ~~-~~3.4 index ~~,~~representing ENSO variability, were generated using four different methods~~to which surface~~. Surface temperature data from the GFDL CM2.1 ~~was applied as a pseudoproxy~~were used as pseudoproxies for these reconstruction methods. As well as sensitivity of the reconstruction to the method, the experiments tested the sensitivity of the reconstruction to the number of non-stationary pseudoproxies and the location of these proxies.

~~ENSO reconstructions in the pseudoproxy experiments were not sensitive to non-stationary teleconnections~~ We find that non-stationarities can act to degrade the skill of ENSO variance reconstructions. However, when global, ~~uniformly-spaced networks of~~ randomly-spaced networks (assuming a minimum of approximately 20 proxies~~were employed~~) were employed, the resulting pseudoproxy ENSO reconstructions were not sensitive to non-stationary teleconnections. Neglecting proxies from ENSO's center-of-action still produced skillful reconstructions, but ~~the chance of generating a skillful reconstruction decreased. Reconstruction methods that utilized raw time series were the most sensitive~~ with a lower likelihood. Different reconstruction methods exhibited varying sensitivities to non-stationary ~~teleconnections, while calculating the running variance of pseudoproxiesfirst, appeared to improve~~ pseudoproxies, which affected the robustness of the resulting reconstructions. The results suggest that caution should be taken when developing reconstructions using proxies from a single teleconnected region, or those that use less than 20 source proxies.

1 Introduction

Reconstructions of the Earth's climate prior to instrumental records are necessary for providing context for anthropogenic climate change, and to provide insight into climate variability on time scales longer than instrumental records allow. Climate proxies are biotic or chemical analogues that have a sensitivity to some aspect of the climate, for example ~~,~~ - oxygen isotope ratios in coral growth rings contain information on temperature and precipitation (Pfeiffer et al., 2004). ~~Thus, these proxies are the essential tool for creating paleoclimate reconstructions.~~ However, high quality proxies can be sparse and difficult to find (McGregor et al., 2010; Neukom and Gergis, 2012), limiting the amount of information that can be inferred about the climate.

One region where information from paleoclimate proxies is limited is the central and eastern tropical Pacific Ocean. This area can be described as the centre-of-action of the El Niño-Southern Oscillation (ENSO), which is the most important regulator of interannual climate variability, globally. ENSO involves changes in eastern equatorial Pacific sea surface temperature (SST) and an associated swing in precipitation and pressure anomalies across the tropical Pacific Ocean. While its most noticeable effects are in the tropical Pacific region, it also induces downstream effects, influencing climate variability in many parts of the world via teleconnections (e.g. Power et al., 1998; Brönnimann et al., 2006; Liu et al., 2013; Ding et al., 2014).

Due to the global reach of ENSO, understanding its behaviour is of great societal and economic importance (Solow et al., 1998; McPhaden et al., 2006). There are still uncertainties about past ENSO (Gergis and Fowler, 2009) and whether ENSO behaviour will change in response to future climate change ~~(Collins et al., 2010; Vecchi and Wittenberg, 2010; Yeh et al., 2014)~~ (Collins et al., 2010; Vecchi a One reason for this is that the instrumental record is too short (∼ 150 years) to measure long term changes in ENSO and its teleconnections (Wittenberg, 2009; Gergis et al., 2006, references therein). Modelling suggests that five centuries of data may be required to understand the full range of natural ENSO variability (Wittenberg, 2009). Thus, climate proxy reconstructions of past fluctuations in ENSO are an essential tool in determining the full range of natural ENSO variability.

As previously described, the centre-of-action of ENSO is largely devoid of long, continuous, high-quality paleoclimate proxy records (Wilson et al., 2010). Tropical corals are the dominant proxy type in this region, and are known to provide very skilful reconstructions of the surrounding SSTs and ENSO. However, the addition of non-climatic noise to these proxies also complicates the estimation of the significance of changing in past ENSO variability (Russon et al., 2014, 2015) , as does their limited life span ~~results in records that~~ (i.e., records are on average about 50 yrs in length, with the longest records less than two centuries~~(Cobb et al., 2013; Neukom and Gergis, 2012)~~) (Cobb et al., 2013; Li et al., 2013; Neukom and Gergis, 2012) . This has motivated the use of paleoclimate proxies from single or multiple regions that are teleconnected with ENSO for the generation of reconstructions. For example, ENSO reconstructions have been developed using paleoclimate proxies from the south-west US and northern Mexico (D'Arrigo et al., 2005), northern New Zealand (Fowler, 2008) and using multiple proxies from locations in the tropical and subtropical Pacific outside ENSO's centre-of-action ~~(Braganza et al., 2009; Wilson et al., 2010)~~ (Braganza et al., 2009; Cobb et al., 2013; Wilson et a Multi-proxy reconstructions are generally considered to be more robust and more likely to contain a larger climate signal to local noise ratio (Mann et al., 1998; Gergis and Fowler, 2009).

There are several issues when using teleconnected proxies for paleoclimate reconstructions. Teleconnections may be non-linear in nature, for example, responding to El Niño events much more strongly than La Niña events (Hoerling et al., 1997). If this is not detected and accounted for in the reconstruction, ENSO variability and amplitude may be misrepresented (McGregor et al., 2013). However, perhaps an equally important issue, is the variability of the teleconnection itself. ENSO reconstructions exploiting teleconnected locations implicitly assume that the teleconnected relationship does not vary significantly in time – that it is stationary. However, it is often difficult or impossible to assess stationarity due to the brevity of the instrumental records (Gallant et al., 2013), causing many to skip this check altogether, noting it as an assumption.

However, significant changes in the relationship between ENSO and the climates of remote, teleconnected locations have been detected in models (Coats et al., 2013; Gallant et al., 2013), instrumental observations (López-Parages and Rodríguez-Fonseca, 2012; Gallant et al., 2013) and paleoclimate data (Hendy et al., 2003; Rimbu et al., 2003; Timm, 2005). If these teleconnections were changed by some dynamical regime rather than through stochastic influence (e.g. random weather events), the relationship should not be considered as stationary. While these dynamical changes could be related to external climate forcing, such as with anthropogenic climate change (Müller and Roeckner, 2008; Herceg Bulić et al., 2011), there is evidence that they also change with internal climate forcing. For example, significant changes in teleconnections on near-centennial time scales are apparent in model simulations forced by internal dynamics alone (Gallant et al., 2013).

The changes to teleconnections via internal dynamics will result from either changes to ENSO itself~~(i.e., changes in the spatial structure of the SST anomalies)~~, or from nonlinear interactions with other regulators of climate variability. An example of the latter is the Southern Annular Mode, which is thought to affect the magnitude of south Pacific ENSO teleconnections (Fogt et al., 2011). The evidence suggests that this occurs on time scales around 30 years or longer. Using running correlations as a statistical descriptor of the relationship between ENSO and a remote climate variable, several studies highlighted that running correlations employing 11–25 year windows of data exhibit large, stochastic variability only (Gershunov et al., 2001; Sterl et al., 2007; van Oldenborgh and Burgers, 2005). However, a study using longer windows of data spanning 31–71 years (Gallant et al., 2013), found that stochastic processes could not explain the changes in observed and modelled running correlations in a significant number of locations in Australasia. Similar results are also found using model simulations (Coats et al., 2013; Gallant et al., 2013). Thus, there are numerous locations that display changes in ENSO's teleconnections that can be classified as "non-stationary" and thus, are thought to be due to dynamical processes. This places increasing stress on the assumption that teleconnections are stationary. ~~Further to this, it~~ This raises the question as to whether non-stationarities have an appreciable influence on the robustness of past paleoclimate reconstructions.

This study examines if and when non-stationary teleconnections degrade the skill of multi-proxy reconstructions of ENSO ~~variability~~ variance by employing a series of pseudoproxy experiments from a fully coupled global climate model (GCM). The robustness of ENSO variance changes (Russon et al., 2014, 2015) is not examined in this paper. The experiments test how reconstruction skill varies with different proxy network locations and sizes. The sensitivity of the results to the reconstruction method is also tested. The model and the data used for these experiments is described in Sect. 2 and the methods are described in Sect. 3. The experimental outcomes are presented in Sect. 4, discussed in Sect. 5 and conclusions are provided in Sect. 6.

2 Model data

This study uses 500 years of a pre-industrial control run of the Geophysical Fluid Dynamics Laboratory Coupled Model 2.1 (GFDL CM2.1) for all pseudoproxy experiments, which are described in detail in Sect. 3. ENSO is represented using the Niño 3.4 index, calculated from the model as the area average of SST anomalies from the central Pacific region (5° S–5° N, 190°–240° E). In the GFDL CM2.1 simulations, the monthly variations in the Niño 3.4

index very closely correspond to the variations of the first Empirical Orthogonal Function (EOF) of tropical Pacific SSTs, demonstrating that the Niño 3.4 index accurately represents ENSO variability in the model (Wittenberg et al., 2006).

Using climate data directly from GCMs is ideal for the evaluation of reconstruction methods (Zorita et al., 2003; Lee et al., 2008; von Storch et al., 2009) because models can provide the long time series necessary to robustly assess multidecadal to near-centennial scale variability in teleconnections (Wittenberg, 2009). The ENSO indices can be calculated directly from the model, representing a "true" Niño 3.4 index for the reconstructed indices to be compared to. This allows the skill of reconstructions to be compared and their sensitivities to be studied.

The GFDL CM2.1 simulation fixes all external climate forcings at 1860 levels. Thus, any changes to ENSO teleconnections will be the product of internal variability only. The model is fully coupled and comprises of the Ocean Model 3.1 (OM3.1), Atmospheric Model 2.1 (AM2.1), Land Model 2.1 (LM2.1), and the GFDL Sea Ice Simulator (SIS). The OM3.1 resolution is 1° latitude by longitude with increasing resolution equatorward of 30°, with 50 vertical layers and a tripolar grid (for more information see Griffies et al., 2005). The AM2.1 and LM2.1 resolution is 2° latitude by 2.5° longitude with 24 vertical levels in AM2.1. For more information on AM2.1 and LM2.1, see Delworth et al. (2006).

The GFDL CM2.1 was selected due to its realistic representation of ENSO characteristics (Wittenberg, 2009, references therein). The seasonal SST structure and ENSO evolution is well represented when compared to observations (Wittenberg et al., 2006; Joseph and Nigam, 2006), while also matching their power spectra (Wittenberg et al., 2006; Lin, 2007). The representation of the strength of local teleconnections in the model, Fig. 1b, shows that the regional responses of surface temperature (TS) and the Niño 3.4 index (shading) are quite similar to the observations (contours). Note that hereafter "TS" refers to SST temperatures over model ocean points and land surface temperatures over model land points. Hence, ENSO in the GFDL CM2.1 is imposing downstream effects, i.e. teleconnections, that are broadly consistent with the observations, even if the strength of the connection is not as is observed (Wang et al., 2012). It has also been shown that the model teleconnections, represented by correlations in 31 year windows between grid points and the Niño 3.4 index generated from the model, do ~~exhibit variability between periods and~~ change over time, and differ compared to correlations calculated over the entire period (Fig. 1a, Wittenberg, 2012). There is significant variation in teleconnection strength (i.e. the range of possible correlations) when using shorter windows of data compared to those of the entire data set.

It has been noted that the strengths, temporal and spatial structures of localised ENSO teleconnections can be poorly represented in GCMs (Joseph and Nigam, 2006; Rowell, 2013; Gallant et al., 2013). This is also seen in CM2.1, as there are teleconnections that are poorly represented at the local level, particularly on the “edges” of the main teleconnections regions (e.g. on the coast of Australia and North America). This is due to inaccuracies in the representation of the mean climate, annual cycle, ENSO, and the other modes of climate variability that are influenced by, or which influence, ENSO, such as the Southern Annual Mode (Delworth et al., 2006). While this limits the conclusions that can be drawn about real-world teleconnections, it still allows for an examination of reconstructions and the associated influence of the non-stationarity of teleconnections, internal to the GCM.

As ENSO events are generally synchronised to the seasonal cycle, the modelled TS was converted to ~~June–July~~ July–June averages to capture ENSO event initiation and termination within one year (Rasmusson and Carpenter, 1982; Tziperman et al., 1997). This has the added benefit of reducing 500 years of monthly TS data (6000 values) to 499 annual values, ~~minimising the computational cost and~~ matching the resolution of the majority of ENSO proxies. The 499 year mean was removed from the dataset and the grid point time series were then linearly detrended by calculating the residuals from a line-of-best fit using linear regression, to remove long-term trends such as model drift. This modified TS dataset is used for all calculations and experiments in this study. Modelled precipitation, only briefly discussed in Sect. 4, was subjected to the same processing prior to any calculations.

3 Methods

This section describes how the model data is used as a substitute for climate proxies and ~~are~~ is selected for multi-proxy reconstructions. Non-stationarity in this paper is defined in Sect. 3.2, and the paleoproxy reconstruction methods tested will be described in Sect. 3.3.

3.1 Pseudoproxy generation

The model TS and precipitation data were used to represent the climate proxies for all reconstructions. These data are commonly referred to as pseudoproxies and represent a “perfect” proxy, free of non-climatic noise (von Storch et al., 2009). ~~Unlike Lee et al. (2008) , these~~ The pseudoproxies are not degraded by adding noise ~~(which would add realism)~~(e.g. Lee et al., 2008) , as the effects of noise on the reconstructions are not in the scope of this study. Pseudoproxies are randomly selected from a subset of the globe, determined by several conditions, depending on the experiment. The most basic condition, present in all experiments, is that the absolute correlation between the model grid point and the Niño 3.4 index is above 0.3 in the calibration window. This threshold is an arbitrary criterion that is simply there to ensure the pseudoproxies ~~represent ENSO to some extent, making them at least partly relevant for reconstructing the ENSOsignal~~at least partially represent ENSO. The calibration window is the time period where relationships between the TS grid points and the model Niño 3.4 index are established. It is entrusted to the reconstruction methods to enhance the signal to noise ratio.

Networks of three to 70 pseudoproxies were used so that the effect of increasing network size could be examined. The same pseudoproxy was not used in the same network more than once, but could be used in multiple networks. ~~One thousand random networks were selected and used to~~ To produce reconstructions of the model Niño ~~-~~3.4 index variance, 1000 random networks were selected per network size, calibration window length, and calibration window position. The randomised selection process over a large number of grid points means that there is only a very small chance that a network would be replicated within 1000 iterations.

The correlation ~~at~~ between ENSO and each grid point time series (i.e. Niño3.4 & TS) over the whole time period (499 years) ~~and ENSO~~ is assumed to represent the true teleconnection strength~~, as its use for calibrating the proxies should result in more accurate reconstructions~~. In reality ~~,~~ however, information is limited to the observational record. As such, calibration can only occur during a relatively brief period, which we expect to result in reconstructions that are not as accurate as they potentially could be. To assess the effects of the use of different calibration windows, we carry out three versions of each experiment.

- The first version represents the scenario where all pseudoproxies with a good correlation, defined as $|r| \geq 0.3$ ~~,~~ over the whole time period (499 years long), can be used in the reconstructions (Fig. 1b). This can be conceptualised by using Fig. 1a, with this ~~series~~ version corresponding to selecting the areas where $|r| > 0.3$ on the x axis (~~where~~ r is 499 year correlation). Information from the entire time series is available in this scenario, and can be thought of as using a calibration window 499 years long.

- The second version represents the realistic scenario, where calibration information is restricted to within a relatively small window and the long term correlation is unknown, much like the effects of limited instrumental data in reality. This can be thought of selecting the areas where $|r| > 0.3$ on the y axis (~~where~~ r is correlation in the calibration window). This implies that there is a chance that the mean correlation over the whole time series is zero, or perhaps the opposite to the expected sign, and this is when non-stationarities are likely to be the largest problem for reconstructions. This would vary with calibration window, and is reflected in Fig. 2b, d and f, with the narrowing of the percentile lines as the length of the calibration window increases.

- The third version represents a combination of the first two ~~series~~version, selecting the proxies with a good correlation in the calibration window, but also over the whole time period (which would normally be unknown). This is equivalent to the case where a proxy is selected during a calibration period, but also happens to have good correlations outside the window – the ideal proxy. This is represented by the overlapping areas of the first two ~~series~~ versions in Figs. 1a, and 2b, d and f for corresponding window lengths. This scenario uses a small calibration window like the second version of experiments, but uses information from the 499 years of data as an additional more stringent pseudoproxy selection criterion.

The first and third versions of experiments produced substantially better reconstructions than the second version. This was ultimately because using much larger calibration windows and using information about the long term strength of teleconnections results in more robust reconstructions. However, in reality, the generation of paleoclimate reconstructions would apply an assumption equivalent to that of the second version of experiments, which limit the information on teleconnection strength to the calibration period only as they are constrained by the instrumental record. ~~However, our experiments showed that this assumption also produces larger errors in the reconstruction (not shown).~~

~~For the remainder of the paper, we show~~ As the second case is the most realistic case, we mainly focus on the second version of ~~the experiments only, as it represents the most realistic case~~experiments for the remainder of the paper. For each grid-box, the 499 year time series was split into ten calibration windows, of lengths 31, 61 and 91 years to match the running correlations performed previously. The mid-point of the calibration windows were spaced evenly in the 499 year dataset, regardless of the amount of overlap or gap between them. Experiments were repeated for the different calibration window lengths and positions, so that the sensitivity of reconstruction skill to calibration window characteristics could be examined. This resulted in ten thousand reconstructions for each calibration window length, for each experiment. The experiments based on pseudoproxy selection are described in Sect. 4.

3.2 Identifying non-stationarities

This study examines the conditions when non-stationary teleconnections impact the validity of paleoclimate reconstructions. Therefore it is necessary to identify which grid points have non-stationary teleconnections, so that its impact on the reconstruction of ENSO can be assessed. The strength and variability of a location’s relationship with ENSO was measured by calculating the running correlation between the grid point TS or precipitation time series, and the modelled Niño 3.4 index. Running correlations used windows of 31, 61 or 91 years, in order to examine multidecadal scale variations on a number of time scales.

This study uses the same definition of non-stationarity as described in detail in Gallant et al. (2013). Non-stationarity was tested against the null hypothesis that the running correlations from the GFDL CM2.1 were stationary. For this purpose, the running correlations computed from the GFDL CM2.1 were compared to the expected range of variation that the running correlations would exhibit if they were only influenced by random noise (e.g. weather events) at the grid point location~~only~~. A Monte Carlo approach (similar to van Oldenborgh and Burgers, 2005; Sterl et al., 2007; Gallant et al., 2013) was used to generate stochastic simulations of TS and precipitation data at each grid point. The simulated data were constructed to have the same statistical attributes as the TS and precipitation data from the GFDL CM2.1 simulation. One thousand stochastic time series were computed for each grid point in order to determine this range, according to the following equation from Gallant et al. (2013).

$$v(t) = a_0 + a_1 c(t) + \sigma_v \sqrt{1-r^2}[\eta_v(t) + B\eta_v(t-1)] \tag{1}$$

$v(t)$ is the stochastic TS or precipitation time series. The first two terms represent the stationary teleconnection strength, with a_0 and a_1 the regression coefficients between the grid point temperature or precipitation and the Niño 3.4 index $c(t)$. The other terms represent the added noise. A red noise process $\eta_v(t) + B\eta_v(t-1)$, was used and is weighted by the standard deviation σ_v of the local TS or precipitation time series, and the proportion of the regression's unexplained variance $\sqrt{1-r^2}$ (where r is correlation of the local time series to the Niño 3.4 index). The red noise is generated by the sum of Gaussian noise (η_v) and autocorrelation (B) of the TS or precipitation time series at lag of 1 year.

A 95 % confidence interval was generated at each grid point from the stochastic simulations and was used to represent the range of running correlations possible, assuming a teleconnection was stationary. Thus, if a running correlation from the GFDL CM2.1 fell outside the range from the stochastic simulations, it was unlikely to have been influenced by stochastic processes alone. Hence, the teleconnection is defined as non-stationary. However, as a 95 % confidence interval was employed, and assuming independent and identically distributed data, such a test would falsely detect a non-stationarity in around 5 % of the time series. So, to decrease the likelihood of detecting false-positives in the time series of running correlations a grid point was defined as non-stationary only if the model running correlation time series fell outside the 95 % confidence interval more than 10 % of the time, which is double than expected by chance alone. As correlations are bounded, the running correlations were converted to Fisher Z scores using the following equation.

$$Z = 1/2\ln[(1+r)/(1-r)] \qquad (2)$$

Z is the Fisher Z score, while r is the running correlation values.

Figure 2a, c and e shows the number of non-stationary ~~years~~ windows (outside the 95% confidence interval) identified in the TS time series at each grid point for the different running correlation windows. Note that the points classified as non-stationary are denoted by the coloured areas in panels a, c, and e, while white areas indicate stationary teleconnections. There are more non-stationary grid points (N value on plot) with larger running correlation windows, suggesting that non-stochastic influences on teleconnections increase as time scales increase. Of further note is a large non-stationary area in the equatorial Pacific~~,~~ ; given this is the area surrounding our ENSO index it is debatable whether this should be considered as a non-stationarity. Rather, we expect the changing relationship in this surrounding region to be the result of ~~ENSO's~~ complexities of ENSO that may not captured by the simple stochastic model of stationarity. For instance, ENSO displays: (i) significant non-linearities ~~(An and Jin, 2004) and/or~~ in its magnitude (An and Jin, 2004) and duration (Okumura and Deser , 2010) ; (ii) differences in the evolution of events with La Niñas and most small to moderate El Niños having SSTAs (sea surface temperature anomalies) that propagate from east to west, while the SSTA of large El Niño events propagate from west to east (Santoso et al., 2013) ; and (iii) changes in its spatial structure (CP-EP ~~type~~ events) which may be considered different flavours of events rather than ~~non-stationarity~~ non-stationary teleconnections of the event (Gallant et al., 2013; Sterl et al., 2007).

3.3 Reconstruction methods

This study examines the ~~likely~~ potential effects of non-stationarities on multi-proxy reconstructions of the running variance of the Niño 3.4 index (representing the variability of ENSO) using pseudoproxy data. All running variances were calculated using 30 year windows. Four simple, commonly-used multi-proxy reconstruction methods were selected. In some methods, such as composite plus scaling (CPS), there are variants to the technique designed to improve climate proxy reconstructions (Jones et al., 2009). However, the impact of non-stationarity on these will not be examined in this study. The reconstruction methods to be tested are as follows:

3.3.1 Median Running Variance (MRV) method

The MRV method was developed by McGregor et al. (2013) to reconstruct the running variance of paleo-ENSO from climate proxy data. It involves calculating the running variance of each of the normalised (zero mean and unit variance) proxy time series, and then calculating the median of these time series. The selected proxies have a demonstrated link to ENSO, identified by a correlation above the prescribed value, to ensure the resulting median time series contains information about ENSO variability.

3.3.2 Running Variance of Median (RVM) method

This method was also devised by McGregor et al. (2013), as an alternate to the MRV for calculating ENSO running variance. Here, if the constituent pseudoproxy series is negatively correlated to ~~Nino~~Niño 3.4, it is flipped in sign before being used for calculations. Each of the proxy time series are normalised to zero mean and unit variance before the median of the group is calculated. This median time series is then normalised prior to calculating its running variance, which is the RVM reconstruction. Despite only differing in the order of operations with the MRV, this method was included in the study as it uses raw time series data, rather than pre-processed data as for the MRV method.

3.3.3 Composite Plus Scaling (CPS) method

CPS is a common method for reconstructing climate data from climate proxies (Esper et al., 2005; Hegerl et al., 2007; Mann et al., 2007, and references therein). In this study, the CPS described in Esper et al. (2005); Hegerl et al. (2007) is employed. The proxy time series are normalised to zero mean and unit variance and are weighted by their correlation to Niño 3.4, before being summed to form a single time series. After normalising this single time series, running variance is taken to reconstruct ENSO variance, hereafter called “CPS_RV”.

3.3.4 Empirical Orthogonal Function Principal Component (EPC) method

This method, described in detail in Braganza et al. (2009), is based on the ability of Empirical Orthogonal Functions (EOFs) to extract the leading modes of variability from a dataset (Xiao et al., 2014, and references therein). Like the MRV method, the proxy data must have established connections to ENSO to ensure that the common dominant signal is an ENSO signal. The leading EOF is then multiplied by the original pseudoproxies, and summed to produce a principal component (PC) time series that is a reconstruction of the ENSO index. The sign of the leading EOF is flipped, if necessary, to ensure that the resulting PC has a positive correlation with the modelled ENSO. Like the CPS method, the running variance of this normalised PC time series is calculated to produce a reconstruction of ENSO variance (hereafter named “EPC_RV”).

3.4 Reconstruction performance

To measure the skill of the reconstructions, each are quantitatively compared to the running variance of the ENSO index in the model (calculated in Sect. 2) by calculating Pearson correlation coefficients and root-mean-squared error (RMSE). Figure 3 shows that each of these four methods capture the running variance well when the entire dataset is available (and with larger proxy networks). Therefore, these methods can be viewed as effective in performing climate reconstructions of ENSO variance. Using all data, the CPS_RV method performs significantly better than the other methods (to a 1 % level of the two-sample Kolmogorov–Smirnov test and Mann–Whitney U test), while the RVM is the worst performing index.

4 Results

The results of the pseudoproxy experiments are presented in this section. Calibration windows of 31, 61 or 91 years are used to generate the reconstructions, and this window length also corresponds to that used for the running correlation. Only grid points with a good absolute correlation to ENSO (> 0.3) within the given calibration window were used as pseudoproxies. Here we examine the sensitivity of the reconstruction methods to non-stationarities, and the effect of proxy location on reconstruction skill. As stated previously, there will be a focus on the reconstructions produced using grid point TS as the pseudoproxies.

4.1 Proxy location effects

ENSO reconstructions are thought to be affected by the locations of the constituent proxies, with many viewing proxies from within the tropical region with higher regard than those sourced elsewhere. These proxies are closest to the centre-of-action and thus expected to be more skilful. Here we examine the impact of tropical Pacific region proxies on reconstructions by comparing two experiments; RND$_{\text{glb\_ts}}$ which selects n pseudoproxies randomly from the global domain (see Supplement Fig. ~~1~~ S1 for locations), while RND$_{\text{ntrop\_ts}}$ has similar random selection but excluding the tropical region: 10° S to 10° N, 100 to 300° E (RND$_{\text{ntrop\_ts}}$). Note that both experiments do not discriminate between stationary and non-stationary locations in this section.

The reconstruction skill, which is represented by the correlation between the pseudoproxy reconstruction of the Niño 3.4 ~~index from the pseudoproxy grid points~~ running variance and the model Niño 3.4 ~~index~~running variance, of both experiments is presented in Fig. 4. Here, network size n is varied from three to 70 (described in Sect. 3.1) on the x axis of each panel, while rows represent the different sized calibration windows and columns the different reconstruction methods (see Sect. 3.3). Looking at the percentile range (Fig. 4, shading) of the correlations between experiments reveals that the removal of tropical Pacific proxies clearly acts to decrease the skill of the resulting reconstructions.

These differences are most easily highlighted by arbitrarily defining skilful reconstructions by some threshold and calculating what proportion of experiment's reconstructions can be classified as skilful. Here we define skilful reconstructions as those that explain more than half the variance of the model ENSO ~~variability~~ variance (grey line at ~ 0.7 correlation). The proportion of skill metrics for the global $\mathrm{RND}_{\mathrm{glb\_ts}}$ and non-tropical $\mathrm{RND}_{\mathrm{ntrop\_ts}}$ experiments, which are respectively plotted in each panel of Fig. 4 as blue and orange lines, can then be further simplified by focusing on the skill difference between experiments (Fig. 4, black line). The skill difference shows clear calibration window length and reconstruction method differences that will be discussed further in Sect. 4.3, but on average when tropical proxies are not used in reconstructions, the proportion of skilful reconstructions decreases by 14 %. However, even without the tropical proxies, the $\mathrm{RND}_{\mathrm{ntrop\_ts}}$ experiment still produced quite high proportions of skilful reconstructions for larger network sizes ($>$ 20 proxies, 77 %). This implies that although there is a reduction in skill with extra-tropical proxies, non-tropical reconstructions still have a high likelihood of producing skilful reconstructions.

4.2 Effect of non-stationarities

Here we examine the effect of non-stationarities on reconstructions of ENSO in order to understand how they may impact past reconstructions of ENSO ~~variability~~variance. To this end, we compare the results of two experiments; (i) $\mathrm{STAT}_{\mathrm{ntrop\_ts}}$, which selects pseudoproxies from the same region as $\mathrm{RND}_{\mathrm{ntrop\_ts}}$ but only includes pseudoproxies that are considered stationary (see definition in Sect. 3.2), while (ii) $\mathrm{NSTAT}_{\mathrm{ntrop\_ts}}$ selects from the same region, but only the non-stationary pseudoproxies. Thus, here we effectively separate the psuedoproxies of the $\mathrm{RND}_{\mathrm{ntrop\_ts}}$ experiment into stationary and non-stationary subgroups and generate reconstructions from each.

Figure 5 has the same panel layout as Fig. 4, with the green and pink representing stationary ($\mathrm{STAT}_{\mathrm{ntrop\_ts}}$) and non-stationary ($\mathrm{NSTAT}_{\mathrm{ntrop\_ts}}$) experiments respectively. Shading represents the percentile ranges of the reconstruction skill, thick lines indicate the proportions of skilful reconstructions and the thick black line is the difference between the stationary (STAT$_{\text{ntrop\_ts}}$) and non-stationary (NSTAT$_{\text{ntrop\_ts}}$) experiments. ~~In~~ For all calibration window lengths (rows) and reconstruction methods (columns), the stationary experiment has greater skill than the non-stationary experiment, although there is reasonable variation between reconstruction methods and calibration window lengths (this will be discussed in later sections). In some cases, non-stationarities can reduce the proportion of skilful reconstructions by up to 60 % (panel b, black line, $n > 60$), but on average the proportion of skilful reconstructions is reduced by 30 %. Thus, these experiments suggest that extra-tropical non-stationarities act to reduce reconstruction skill.

It is interesting to note that when tropical region non-stationarities are included, they appear to improve reconstruction skill (Supplement Fig. ~~S4~~S3). The majority of the pseudoproxies in the tropical region were found to be highly correlated with ENSO as expected, and to demonstrate very little variation in their correlations to ENSO (not shown), usually less than ~ 0.1 correlation. However, as seen in Fig. 2 many of these proxies are still classified as non-stationary, which may be due to non-linearities or variations in flavour of ENSO events. Thus, regardless of whether they are classified as non-stationary or not, the inclusion of these tropical pseudoproxies acts to improve the skill of the ENSO reconstructions.

~~In regards to why~~ The fact that we see a minimal effect of non-stationarities ~~do not seem to impact the high skill of random pseudoproxy selection of~~ in the randomly selected experiments (see Sect. 4.1~~, we find that~~) may be because the likelihood of selecting non-stationarities is relatively low. For instance, Fig. 6 shows the proportions of non-stationary pseudoproxies in the reconstructions for the RND$_{\text{glb\_ts}}$ experiment with a 31 year long calibration window. It varies with different proxy network sizes, but as expected, the smaller groups have a greater chance of higher proportions of non-stationary proxies. With networks greater than thirty, the most likely proportion is around 14 %, while much more consistent than the smaller groups. Even with very small group sizes ($n = 3$), the chance that all stations are non-stationary is only 0.3 % (red line from Fig. 6). When only using extratropical locations (RND$_{\text{ntrop\_ts}}$), the most likely proportion of non-stationary proxies is around

9 %, with an even lower chance of all constituent proxies being non-stationary. There is also a tendency for more non-stationarities to occur with the use of longer calibration windows (see Fig. 2a), consequently the proportions of non-stationary proxies increase. For example, networks greater than thirty proxies can be up to 25 % non-stationary when using 91 year calibration windows (not shown). Regardless of the increases in non-stationarities with the use of longer calibration windows, these longer windows still produced more skilful reconstructions in the random selection experiments than those with shorter windows ($\mathrm{RND}_{\mathrm{glb\_ts}}$ and $\mathrm{RND}_{\mathrm{ntrop\_ts}}$; Fig. 4). Thus, although non-stationarities have the potential to influence the skill of ENSO reconstructions, this scenario appears unlikely if proxies are selected similar to a globally random manner.

However, if pseudoproxies are selected from regions that ~~have non-stationarities occurring at the same time~~demonstrate co-variability in the running correlation between TS and Niño 3.4 SST anomalies, reconstruction skill is devastated. To this end, an Empirical Orthogonal Function analysis (EOF) was ~~essentially~~ used to “organise” ~~the non-stationarities, resulting in the experiment~~ this co-variability, of which it is expected that non-stationarities are a major part. This is seen in the PNEOF1 experiment shown in Fig. 7. In this experiment the EOF was carried out on the running correlations between TS and Niño 3.4 SST anomalies at each grid point. Pseudoproxy networks were then selected only from those grid points that exhibited a strong relationship with the leading EOF (i.e. the absolute value of the EOF weighting ~~> 0.1~~> 0.01). The spatial map of this leading EOF is shown in panel e, for 31 year window running correlations. The leading EOFs of the longer windows have very similar spatial patterns, with spatial correlations of 0.86 and 0.84 produced respectively, when comparing the 61 and 91 year window length EOF1 spatial patterns (not shown). The leading principal components for each window length are also similar (panel f). The resulting PNEOF1 experiment reconstructions display a large loss in skill when compared to the stationary pseudoproxies in the reconstructions ($\mathrm{STAT}_{\mathrm{ntrop\_ts}}$, dashed lines), with the former having very little likelihood of producing a skilful reconstruction (Fig. 7~~a). This highlights that non-stationarities can significantly affect the skill of reconstructions if there is spatial coherence in the non-stationarities~~a–d). The proportion of non-stationary grid points used in the PNEOF1 reconstructions was small, ranging from 9-15%. However, there was still a substantial loss of skill in these reconstructions even though the majority of grid points were classified as stationary by our statistical definition. This implies that a large and coherent change to the teleconnection exists in that region even if it considered mostly statistically stationary, and that was enough to degrade reconstruction skill. Thus, care should be taken to avoid the scenario where all constituent pseudoproxies ~~of a reconstruction can have non-stationarities occurring at the same times~~used in a reconstruction lie in a region where there are large, coherent variations in teleconnections, even if these variations are considered stationary.

4.3 Pseudoproxy network size and length

As ~~shown~~ discussed previously, the ENSO reconstruction skill is sensitive to the pseudoproxy network size and window length. This is clearly seen in Fig. 8, which displays the reconstruction skill of three different previously presented experiments (RND$_{\text{glb\_ts}}$, RND$_{\text{ntrop\_ts}}$, and NSTAT$_{\text{ntrop\_ts}}$) ~~. In each panel the three colours indicate which calibration window length is used; 31 (blue)~~as three different colours (see legend). Each panel shows the proportions of skilful reconstructions (thick lines) for different reconstruction methods (as titled), ~~61 (green) , or 91 (red)years, while the hatching is the percentile range, and the thick lines are the proportion of skilful reconstructions~~and different calibration window lengths (see inset legend). What is clear in all panels, is that the reconstruction skill generally improves with increasing network size for all experiments, that is regardless of reconstruction method and calibration window length. This is also true when all pseudoproxies in a network are non-stationary (NSTAT$_{\text{ntrop\_ts}}$ experiment), ~~however,~~ although the reconstruction skill generally improves at a slower rate (Fig. 8~~i, j, l~~, red lines). This implies that larger pseudoproxy networks are less affected by non-stationarities, but this is also dependent on the calibration window length (discussed below) and the reconstruction method (discussed in Sect. 4.4). In ~~general, smaller pseudoproxy networks (<5) produce very low proportions of skilful reconstructions (10–40), while those with larger networks the majority of reconstructions become skilful. In~~ fact, when pseudoproxies are randomly selected

(RND$_{\text{glb\_ts}}$ and RND$_{\text{ntrop\_ts}}$), using a minimum of 20 proxies gives a ~~fairly good chance (> 77~~reasonable chance (77% chance on average) that the resulting reconstruction will be skilful (Fig. 8~~a–c, e–g~~).

The calibration window length also has an impact on reconstruction skill and sensitivity to non-stationarities (Fig. 9). For example, using small calibration windows (31 to 91 years) compared to the total number of model years available (499 years) leads to a relative decrease in skill, as indicated by the black 499 year reconstruction being higher in skill than the reconstructions using smaller windows. This decrease of skill ~~would be~~ is potentially due to some information loss in the relative datasets, and not necessarily due to non-stationarities. However, this reduction in skill at the median (thick line) is quite small (~ 0.1 correlation) even at the smallest networks sizes and in the worst performing reconstruction method. Thus, although there is a reduction in skill due to loss of information with smaller calibration window lengths, this is relatively small compared to the possible impacts of non-stationarities (see previous section). Figure 8 also shows that larger windows tend to improve skill, with the larger window lengths consistently having higher proportions of skilful reconstructions in the random selection experiments (RND$_{\text{glb\_ts}}$ and RND$_{\text{ntrop\_ts}}$). Larger windows also appear to generally improve reconstructions in the NSTAT$_{\text{ntrop\_ts}}$ experiment. ~~However, for random proxy selection, longer calibration windows still lead to increases in reconstruction skill, as long as the proxy network is not entirely non-stationary (like in the NSTAT$_{\text{ntrop\_ts}}$ experiment).~~ This increase in skill is not as great as removing non-stationarities from the reconstructions (Fig. 5) or changing the reconstruction method (following section).

4.4 Reconstruction method comparison

All reconstruction methods create skilful reconstructions given sufficiently large calibration windows and proxy network sizes in the random selection experiments RND$_{\text{glb\_ts}}$ and RND$_{\text{ntrop\_ts}}$ (see Figs. 8 and 9). ~~It is noted that the~~ The CPS_RV method performs ~~well~~almost as well as the MRV, although mainly with longer calibration windows and for the random selection experiments (RND$_{\text{glb\_ts}}$ and RND$_{\text{ntrop\_ts}}$, Fig. 8). However, there is a clear distinction in the skill from the MRV method reconstructions compared to the other methods tested when considering the impact of non-stationarities and neglecting tropical pseudoproxies. For instance, when tropical pseudoproxies are not used in experiments, the MRV reconstructions are only marginally affected (Fig. 4c, g and k) implying that the method is not as dependent as other methods on the highly correlated tropical region. This is expected, as the EPC_RV and CPS_RV involve weighting regimes that would favour the highly correlated tropical pseudoproxies (see Sect. 3.3, and references therein). The MRV method has the highest proportion of skilful reconstructions at the lowest network sizes in all other experiments (Fig. 8c), with the clearest differences seen in the $\mathrm{NSTAT}_{\mathrm{ntrop\_ts}}$ experiment (Figs. 5 and 8~~i–l~~), while the percentile range of the MRV method also tends to be the smallest. Both of which, indicate that the MRV method has the lowest sensitivity to non-stationarities. Further to this, in spite of the MRV method being negatively affected in the PNEOF1 experiment (Fig. 7, thick lines), and displaying some sensitivity to calibration window length (~~red line outperforms others~~91yr windows perform better than shorter windows), it produces the highest proportion of skilful reconstructions and is thus still the most robust against non-stationarities.

It is worth noting that although the MRV method shows the most consistently high correlations to ENSO ~~, this high skill is not necessarily reflected in the RMSE (root-mean-square error). The RMSE of the MRV method is still the most consistent however~~ and appears to be the least sensitive to calibration window position (smallest percentile ranges, Supplement Fig. ~~5), but shows somewhat greater error than the other methods in this experiment (RND$_{\mathrm{ntrop\_ts}}$). MRV in the non-stationary experiment (NSTAT$_{\mathrm{ntrop\_ts}}$, Supplement~~ S5), it has the highest RMSE (root-mean-square error). It is well known that all reconstruction methods result in a loss in ENSO variance, and this is clearly shown in Fig. 10. In Fig. ~~S6; PNEOF1, not shown) have similar RMSE values to other methods, likely due to the other methods gaining additional errors due to increased non-stationarities. Upon further inspection it is clear that the higher correlations of the MRV method are offset by the resulting running variance time series being much more damped than those of the other methods, which explains the high RMSE error. This can be seen in Supplement Fig. S7, where the variance~~

~~is taken of the reconstructions instead of the correlations like in previous analyses. The MRV results clearly show much lower variance than all the other methods~~ 10a–d, we can see that all reconstructions underestimate the model Niño 3.4 running variance (black line). However, this figure also shows that this variance loss is exaggerated with the MRV method (panels c, g~~and k), particularly at larger pseudoproxy network sizes, whilst the variance of other methods remain relatively high with increasing network size. Due to the nature of the other methods~~), and this is also seen in Supplement Fig. S6, particularly at the larger network sizes. It is this variance loss that leads to the high RMSE of the MRV method. Other methods do not suffer as much from this variance loss as they are normalised after the reconstruction but prior to the calculation of the running variance (see Sect. 3.3)~~, while the MRV is not~~. As the MRV utilises running variances from the beginning, it unable to be normalised. Thus, while the MRV reproduces ENSO variance with the highest correlation skill, the MRV method ~~may require~~ requires re-scaling to better match the magnitude of the variance changes.

~~Given that the RVM and MRV methods are only different in order of operations (see Sect. 3.3)their large differences in reconstruction skill suggest that using the median, rather than weighting the individual source time series, plays little role in the robustness of the MRV method. As McGregor et al. (2013) identified, taking running variances first, which are positive definite (see Sect. 3.3), means that the MRV method is not susceptible to signal cancellation like the other methods including the RVM. Thus, we suggest that the MRV method is robust against non-stationarities because they act much like dating errors and lead to signal cancellation.This is supported by Fig. **??**, where a few examples of reconstructions are plotted alongside the standard deviation of their source pseudoproxies' running correlation to model ENSO (see McGregor et al., 2013) .These plots suggest that when~~ In order to compensate for the variance loss of each reconstruction (Fig. 10.a–d), we rescale each method's resulting running variance time series (Fig. 10.e–h). Rescaling the running variance time series was carried out using the average (calculated over 1000 reconstructions) regression between the reconstructions and the modelled Niño 3.4 running variance within the calibration window. When the MRV (panel 10.c) is scaled to form the SMRV (panel 10.g), there is a ~~lot of variability in the correlations between the source pseudoproxies and ENSO, the reconstruction variance tends to be low (and vice-versa), which can be seen in the red highlighted areas. This supports the idea that non-stationarities act to cancel the running~~ jump in reconstruction variance ~~signal much like a dating error. Further to this ,~~ (grey shading), such that the ~~regressions of these individual time series also show the MRV's difference to other methods, with a much smaller regression slope −0.79 for MRV, compared to −2.28, −1.99 and −2.32 for the RVM, CPSSUBSCRIPTNBRV and EPCSUBSCRIPTNBRV methods, respectively (out of the statistically significant reconstructions). Thus, there is evidence that~~ modelled Niño 3.4 index running variance is now encompassed by the grey shading. Using this simple scaling technique, we see a large reduction in the RMSE (see Supplement Fig. S7) - up to a 0.1 reduction in the median (Supplement Fig. S7, cyan lines) and no changes in the correlation (not shown). In fact, it is noteworthy that on average the scaled MRV has the smallest RMSE (significant to the ~~MRV method is less prone to variance losses when there is high variability amongst the source proxies, and hence it is less susceptible to signal cancellation in proxies~~99% level via a two sample t test) of all reconstruction methods.

4.5 Precipitation pseudoproxies

Although not the focus ~~on~~ of this paper, precipitation was also examined for all experiments. Precipitation based reconstructions showed more variation in skill than TS and required larger network sizes for the same skill (see Supplement Fig. S2), but otherwise had similar tendencies as temperature outlined above. However, there was one key difference in precipitation – NSTAT$_{\text{glb\_pr}}$ (~~Supplement Fig. S3~~red lines) produced less skilful reconstructions than RND$_{\text{glb\_pr}}$ (~~Supplement Fig. S2)~~ grey lines) as we would expect. This is likely due to the absence of a large spatially coherent region of correlations in the tropical Pacific Ocean (~~see~~ compare tropical areas in Supplement Fig. ~~S1e). Generally, there is also greater variability in skill across calibration windows than in temperature (~~S1b and Supplement Fig. ~~4, blue shading), leading to wider shaded areas in the EPCSUBSCRIPTNBRV and~~ S1e). The CPS_RV ~~methods, but not~~

~~much change for the MRV and RVM methods. In the precipitation $\mathrm{RND}_{\mathrm{glb\_pr}}$ experiment (Supplement Fig. S2), the CPSSUBSCRIPTNBRV method is generally unskilful, with the worst 5of reconstructions (blue shading) displaying correlations below zero with network sizes below 10 proxies~~method also generally outperforms the other methods, except for the non-stationary experiment $\mathrm{NSTAT}_{\mathrm{glb\_pr}}$, where the MRV appears to be superior to all methods. The RVM method appears to perform slightly better with precipitation than temperature ~~in panels d, and h, with not much difference in panel l~~(Supplement Fig. S2, mainly at longer calibration windows), which is consistent with the findings of McGregor et al. (2013).

5 Discussion

Non-stationary relationships between the modelled Niño 3.4 index and regional temperature and precipitation were detected in the GFDL CM2.1 model. Our results demonstrate that non-stationarities between ENSO and regional climates can occur in many regions around the globe, which extends previous work of Gallant et al. (2013), who found significant non-stationary areas in the Australasian region in both modelling and observations. Like in Gallant et al. (2013), our work shows non-stationarities exist in climate models globally on time scales longer than approximately 30 years, demonstrating their occurrence at low frequencies. This is in contrast to van Oldenborgh and Burgers (2005) and Sterl et al. (2007), who examined non-stationarities at higher frequencies and found no detectable evidence for them in the observations using running correlation windows of around 20 years. The fact that these non-stationarities are found in a pre-industrial control simulation shows that this low frequency variability can arise from unforced, internal climate variability, adding further evidence that this low frequency variability is an inherent part of the climate system.

Identifying what causes the occurrence of non-stationarities in ENSO teleconnections is not within the scope of this study. However, Wittenberg (2009) showed substantial changes to the behaviour of ENSO on similar time scales to those identified here in a 2000 year simulation using the GFDL CM2.1. Wittenberg (2009) discussed that such changes to ENSO behaviour could conceivably alter the teleconnections between ENSO and local climate ~~. We note that although we use the same model as in the Wittenberg (2009) study, the results are unlikely to be a product of the model configurationgiven that Gallant et al. (2013) identified non-stationarities in three different GCMs~~and that these changes may not be represented in the historical record. Gallant et al. (2013) identified non-stationarities in three different GCMs. It is noted that while numerous models display non-stationarities, their regional existence may vary depending on the model used (Coats et al., 2013) . We do not expect our evaluation of various different reconstruction methods performance in the presence of non-stationarities to be affected by model configuration, however we intend to examine this in future research.

In this study, the pseudoproxy approach in the ~~virtual~~ reality of the GFDL CM2.1 pre-industrial control simulations avoids the problems of non-climate related noise that is inherent to real-world paleoclimate proxies, allowing us to focus on the sensitivity of reconstructions to the occurrence of non-stationarities alone. However, in reality non-climate related sources of noise in paleoclimate proxies will confound, and likely degrade, reconstruction skill to a greater extent than examined here. Thus, our finding that a network size of > 20 will minimise the effects non-stationarities on reconstruction skill is likely an underestimate of minimum network size for a real-world reconstruction. The compounding effects of noise and non-stationarities on the reconstruction method and hence, a reconstruction, should be the focus of future research efforts in this area.

All reconstruction methods examined generate skilful reconstructions when utilising globally random source proxy selection, given sufficiently large calibration windows and proxy network sizes. Therefore, the results presented here highlight a case for considering the influence of non-stationarities on real-world reconstructions ~~,~~ and their underlying methods, which generally employ small proxy networks. The influence of the choice of method on the reconstruction and its sensitivity to non-stationarities was stark. The non-stationarities and reconstruction method usually had a greater influence on reconstruction skill than the calibration window length. In the best-case scenario (i.e. long calibration window and large proxy network), the CPS_RV method had the greatest skill. In less-than-ideal conditions (e.g. small calibration windows or proxy networks), the MRV method clearly excelled, and even managed to produce a high proportion of skilful reconstructions given only pseudo-proxies considered non-stationary (Fig. 5). However, ~~note that the performance of these methods is likely to depend on the variable being reconstructed~~the unscaled MRV method showed poor RMSE performance, meaning that it can only be used to provide useful information on the relative changes in ENSO variance. We also note that the large difference between the MRV and RVM experiments (Figs. 3 and 9) is contradictory to the results in Fig. 4 of McGregor et al. (2013). However, these differences were due to the 10 year low-pass filter used in McGregor et al. (2013), whereas in this study, the data was unfiltered. Consequently, the RVM was found to be sensitive to the low-pass filtering while the MRV was insensitive (results not shown).

For reconstructions of large-scale phenomena like ENSO, ~~multi-proxy~~ larger more globally diverse networks will produce more informative reconstructions ~~because the larger networks contain more information, including spatial information, compared to single site~~ compared to those derived from smaller regions or single sites (Mann, 2002; Lee et al., 2008; von Storch et al., 2009; McGregor et al., 2013). The experiments conducted here support this hypothesis, as the proportions of skilful reconstructions increase as the number of source proxies increase for almost all reconstruction methods and calibration window lengths (Figs. 8 and 5). Our work further shows that large ~~, multi-proxy~~ networks also reduce errors relating to non-stationarity of teleconnections, which further supports their employment (Fig. 5). However, this skill improvement is affected by the degree of non-stationarity and teleconnection co-variability present in the reconstructions, with non-stationary proxy networks ($\text{NSTAT}_{\text{ntrop\_ts}}$, Fig. 8~~i–l~~, red lines) and "organised" ~~non-stationarities~~ teleconnection co-variability (PNEOF1, Fig. 7a–d) reducing the degree of improvement in skill with increasing network size. Thus, where increasing network size would usually improve the reconstruction, non-stationarities and spatial coherence in variations in teleconnection strength can substantially temper this improvement. In extreme cases, where proxies are selected from ~~areas with spatially coherent non-stationarities~~ co-varying areas (PNEOF1, Fig. 7), reconstruction skill may show no improvement with larger proxy networks. This further stresses the importance of ensuring that all constituent proxies utilised in a ~~-~~reconstruction are not affected by ~~the same non-stationarities~~co-varying teleconnections. This is more likely achieved in spatially diverse, large multi-proxy networks.

The results of this study further emphasise the need for more paleoclimate proxies to be available for multi-proxy climate reconstructions. Given the skilful reconstructions in ENSO variance that can be produced by neglecting pseudoproxies from the centre of action ~~,~~ as shown here, the utilisation of data solely from the eastern equatorial Pacific appears unnecessary. In fact, these results utilising globally random proxy selection support the development of paleoclimate proxies from a wide range of global locations. Furthermore, developing an understanding of the teleconnections and their underlying mechanisms around the globe will assist with selection of paleoclimate proxy locations that are unlikely to be affected by ~~the same non-stationarity~~teleconnection co-variability.

6 Conclusions

We have demonstrated that non-stationarities in ENSO teleconnected proxies can significantly reduce reconstruction skill, and that this is dependent on proxy location, multi-proxy network size, and reconstruction method. These results ~~assume that the model data is a realistic representation of the~~ make the implicit assumption that the modelled co-variability of the non-stationarities and relative proportions of non-stationary areas to stationary areas ~~, which have~~ are realistic, which has not been explicitly tested here. Ultimately, our results show that non-stationarities are unlikely to significantly affect reconstruction skill for larger, globally selected, multi-proxy networks (>20 proxies). Non-stationarities will deteriorate reconstructions if the entire network exhibits non-stationarities, but this is highly unlikely (<0.3 %) for large networks (>20 proxies), which can be considered globally distributed. However, the results suggest caution when developing reconstructions using single site proxies or multiple proxies from the same teleconnected region, as there is a reasonable chance this would lead to an unskilful reconstruction if there are no other sources of information. Thus, using multiple teleconnected regions minimises any effects of non-stationarities for all methods tested.

Reconstruction methods that ~~allow for signal cancellation when combining proxies (i.e. those that~~ operate on the raw time series data (weighting the proxy time series directly) are most sensitive to non-stationarities (RVM, EPC_RV and CPS_RV methods), while the method utilising the running variance time series (MRV method) is the most robust against non-stationarities. However, these were the only methods tested, and there are many ~~various~~ reconstruction methods in the literature (Jones et al., 2009; Wilson et al., 2010) that should be tested in future research. Neglecting proxies from ENSO's center-of-action still allows for skilful reconstructions to be made, but their inclusion reduces the chance of producing particularly poor reconstructions even if non-stationarities are present.

~~With the short instrumental record, detecting the presence of non-stationarities in teleconnections may be difficult. However, we have shown using a fully coupled GCM that for larger multi-proxy networks selected over broad areas, non-stationary teleconnections are unlikely to affect reconstruction skill. Non-stationarities will deteriorate reconstructions if the entire network exhibits non-stationarities, but this is highly unlikely (< 0.3) for large networks (> 20 proxies), which can be considered globally distributed. As such, we advise caution when using small multi-proxy networks and where the proxies are located within very few teleconnected regions. Although not examined in this paper, our results suggest that teleconnected single-proxy reconstructions would be much more prone to loss of reconstruction skill in the presence of non-stationarities when compared to multi-proxy reconstructions. Thus, we do not advocate their use for reconstructing large-scale climatic processes.~~ Further research would involve examining the organisation of non-stationarities and co-varying teleconnections in more detail, exploring the use of running variance on proxy time series as pre-processing, or evaluating how robust other reconstruction methods are against non-stationary teleconnections.

The Supplement related to this article is available online at doi:10.5194/cpd-0-1-2015-supplement.

Acknowledgements. This research was supported by the Australian Research Council Centre of Excellence for Climate System Science (CE110001028) and the University of New South Wales. Shayne McGregor and Ailie J. E. Gallant are supported by Australian Research Council Discovery Early Career Researcher Awards DE130100663 and DE150101297. The authors would like to thank ~~Associate Professor Katrin Meissner for her~~ our reviewers for their valuable comments on the manuscript. The leading author would like to thank Shayne McGregor for his guidance and contributions to this study.

References

An, S.-I. and Jin, F.-F.: Nonlinearity and asymmetry of ENSO, J. Climate, 17, 2399–2412, 2004.

Braganza, K., Gergis, J. L., Power, S. B., Risbey, J. S., and Fowler, A. M.: A multiproxy index of the El Niño Southern Oscillation, A.D. 1525–1982, J. Geophys. Res., 114, doi:10.1029/2008JD010896, 2009.

Brönnimann, S., Xoplaki, E., Casty, C., Pauling, A., and Luterbacher, J.: ENSO influence on Europe during the last centuries, Clim. Dynam., 28, 181–197, 2006.

Coats, S., Smerdon, J. E., Cook, B. I., and Seager, R.: Stationarity of the tropical pacific teleconnection to North America in CMIP5/PMIP3 model simulations, Geophys. Res. Lett., 40, 4927–4932, 2013.

Cobb, K. M., Westphal, N., Sayani, H. R., Watson, J. T., Di Lorenzo, E., Cheng, H., Edwards, R. L., and Charles, C. D.: Highly variable El Niño-Southern Oscillation throughout the Holocene, Science, 339, 67–70, 2013.

Collins, M., An, S.-I., Cai, W., Ganachaud, A., Guilyardi, E., Jin, F.-F., Jochum, M., Lengaigne, M., Power, S., Timmermann, A., Vecchi, G., and Wittenberg, A.: The impact of global warming on the tropical Pacific Ocean and El Niño, Nat. Geosci., 3, 391–397, 2010.

D'Arrigo, R., Cook, E. R., Wilson, R. J., Allan, R., and Mann, M. E.: On the variability of ENSO over the past six centuries, Geophys. Res. Lett., 32, doi:10.1029/2004GL022055, 2005.

Delworth, T. L., Broccoli, A. J., Rosati, A., Stouffer, R. J., Balaji, V., Beesley, J. A., Cooke, W. F., Dixon, K. W., Dunne, J., Dunne, K. A., Durachta, J. W., Findell, K. L., Ginoux, P., Gnanadesikan, A., Gordon, C. T., Griffies, S. M., Gudgel, R., Harrison, M. J., Held, I. M., Hemler, R. S., Horowitz, L. W., Klein, S. A., Knutson, T. R., Kushner, P. J., Langenhorst, A. R., Lee, H.-C., Lin, S.-J., Lu, J., Malyshev, S. L., Milly, P. C. D., Ramaswamy, V., Russell, J., Schwarzkopf, M. D., Shevliakova, E., Sirutis, J. J., Spelman, M. J., Stern, W. F., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, F., and Zhang, R.: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics, J. Climate, 19, 643–674, 2006.

Ding, Q., Wallace, J. M., Battisti, D. S., Steig, E. J., Gallant, A. J. E., Kim, H.-J., and Geng, L.: Tropical forcing of the recent rapid Arctic warming in northeastern Canada and Greenland, Nature, 509, 209–212, 2014.

Esper, J., Frank, D. C., Wilson, R. J. S., and Briffa, K. R.: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium, Geophys. Res. Lett., 32, doi:10.1029/2004GL021236, 2005.

Fogt, R. L., Bromwich, D. H., and Hines, K. M.: Understanding the SAM influence on the South Pacific ENSO teleconnection, Clim. Dynam., 36, 1555–1576, 2011.

Fowler, A. M.: ENSO history recorded in Agathis australis (kauri) tree rings. Part B: 423 ears of ENSO robustness, Int. J. Climatol., 28, 21–35, 2008.

Gallant, A. J. E., Phipps, S. J., Karoly, D. J., Mullan, A. B., and Lorrey, A. M.: Nonstationary Australasian teleconnections and implications for paleoclimate reconstructions, J. Climate, 26, 8827–8849, 2013.

Gergis, J., Braganza, K., Fowler, A. M., Mooney, S., and Risbey, J. S.: Reconstructing El Niño-Southern Oscillation (ENSO) from high-resolution palaeoarchives, J. Quaternary Sci., 21, 707–722, 2006.

Gergis, J. L. and Fowler, A. M.: A history of ENSO events since A.D. 1525: implications for future climate change, Climatic Change, 92, 343–387, 2009.

Gershunov, A., Schneider, N., and Barnett, T.: Low-frequency modulation of the ENSO-Indian monsoon rainfall relationship: signal or noise?, J. Climate, 14, 2486–2492, 2001.

GISTEMP-Team: GISS Surface Temperature Analysis (GISTEMP), available at: http://data.giss.nasa.gov/gistemp/, last access: 18 March 2015.

Griffies, S. M., Gnanadesikan, A., Dixon, K. W., Dunne, J. P., Gerdes, R., Harrison, M. J., Rosati, A., Russell, J. L., Samuels, B. L., Spelman, M. J., Winton, M., and Zhang, R.: Formulation of an ocean model for global climate simulations, Ocean Sci., 1, 45–79, doi:10.5194/os-1-45-2005, 2005.

Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Rev. Geophys., 48, rG4004, doi:10.1029/2010RG000345, 2010.

Hegerl, G. C., Crowley, T. J., Allen, M., Hyde, W. T., Pollack, H. N., Smerdon, J., and Zorita, E.: Detection of human influence on a new, validated 1500-year temperature reconstruction, J. Climate, 20, 650–666, 2007.

Hendy, E. J., Gagan, M. K., and Lough, J. M.: Chronological control of coral records using luminescent lines and evidence for non-stationary ENSO teleconnections in northeast Australia, Holocene, 13, 187–199, 2003.

Herceg Bulić, I., Branković, C., and Kucharski, F.: Winter ENSO teleconnections in a warmer climate, Clim. Dynam., 38, 1593–1613, 2011.

Hoerling, M. P., Kumar, A., and Zhong, M.: El ~~Nino~~Niño, La Nina, and the nonlinearity of their teleconnections, J. Climate, 10, 1769–1786, 1997.

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., Luterbacher, J., Wahl, E. R., Zwiers, F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Küttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wolff, E., and Xoplaki, E.: High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects, Holocene, 19, 3–49, 2009.

Joseph, R. and Nigam, S.: ENSO Evolution and teleconnections in IPCC's twentieth-century climate simulations: realistic representation?, J. Climate, 19, 4360–4377, 2006.

Lee, T. C. K., Zwiers, F. W., and Tsao, M.: Evaluation of proxy-based millennial reconstruction methods, Clim. Dynam., 31, 263–281, 2008.

Li, J., Xie, S.-P., Cook, E. R., Morales, M. S., Christie, D. A., Johnson, N. C., Chen, F., D'Arrigo, R., Fowler, A. M., Gou, X., and Fang, K.: El Nino modulations over the past seven centuries, Nature Clim. Change, 3, 822-826, doi:10.1038/nclimate1936, 2013.

Lin, J.-L.: Interdecadal variability of ENSO in 21 IPCC AR4 coupled GCMs, Geophys. Res. Lett., 34, doi:10.1029/2006GL028937, 2007.

Liu, N., Wang, H., Ling, T., and Feng, L.: The influence of ENSO on sea surface temperature variations in the China seas, Acta Oceanol. Sin., 32, 21–29, 2013.

López-Parages, J. and Rodríguez-Fonseca, B.: Multidecadal modulation of El Niño influence on the Euro-Mediterranean rainfall, Geophys. Res. Lett., 39, doi:10.1029/2011GL050049, 2012.

Mann, M. E.: The value of multiple proxies, Science, 297, 1481–1482, 2002.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, 1998.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Robustness of proxy-based climate field reconstruction methods, J. Geophys. Res., 112, doi:10.1029/2006JD008272, 2007.

McGregor, S., Timmermann, A., and Timm, O.: A unified proxy for ENSO and PDO variability since 1650, Clim. Past, 6, 1–17, doi:10.5194/cp-6-1-2010, 2010.

McGregor, S., Timmermann, A., England, M. H., Elison Timm, O., and Wittenberg, A. T.: Inferred changes in El Niño–Southern Oscillation variance over the past six centuries, Clim. Past, 9, 2269–2284, doi:10.5194/cp-9-2269-2013, 2013.

McPhaden, M. J., Zebiak, S. E., and Glantz, M. H.: ENSO as an integrating concept in earth science, Science, 314, 1740–1745, 2006.

McPhaden, M. J., Zhang, X.: Asymmetry in zonal phase propagation of ENSO sea surface temperature anomalies, Geophys. Res. Lett., 36, doi:10.1029/2009GL038774, 2009.

Müller, W. A. and Roeckner, E.: ENSO teleconnections in projections of future climate in ECHAM5/MPI-OM, Clim. Dynam., 31, 533–549, 2008.

Neukom, R. and Gergis, J.: Southern Hemisphere high-resolution palaeoclimate records of the last 2000 years, Holocene, 22, 501–524, 2012.

Okumura, Y. M., and Deser C.: Asymmetry in the Duration of El Niño and La Niña, J. Climate, 23, 5826–5843, doi:10.1175/2010JCLI3592.1, 2010.

Pfeiffer, M., Dullo, W.-C., and Eisenhauer, A.: Variability of the intertropical convergence zone recorded in coral isotopic records from the central Indian Ocean (Chagos Archipelago), Quaternary Res., 61, 245–255, 2004.

Power, S. B., Tseitkin, F., Torok, S., Lavery, B., Dahni, R., and McAvaney, B.: Australian temperature, Australian rainfall and the Southern Oscillation, 1910–1992: coherent variability and recent changes, Aust. Meteorol. Mag., 47, 85–101, 1998.

Rasmusson, E. M. and Carpenter, T. H.: Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño, Mon. Weather Rev., 110, 354–384, 1982.

Rimbu, N., Lohmann, G., Felis, T., and Pätzold, J.: Shift in ENSO teleconnections recorded by a northern red sea coral, J. Climate, 16, 1414–1422, 2003.

Rowell, D. P.: Simulating SST teleconnections to Africa: What is the state of the art?, J. Climate, 26, 5397–5418, 2013.

Russon, T., Tudhope, A. W., Hegerl, G., Collins, M., and Schurer, A.: Assessing the significance of changes in ENSO amplitude using variance metrics, J. Climate, 27, 4911–4922, doi:10.1175/JCLID-13-00077.1, 2014.

Russon, T., Tudhope, A. W., Collins, M., and Hegerl, G. C.: Inferring changes in ENSO amplitude from the variance of proxy records, Geophys. Res. Lett., 42, 1197–1204, doi:10.1002/2014GL062331, 2015.

Santoso A., McGregor S., Jin, F.-F., Cai, W., England, M. H., An, S.-I., McPhaden M. J., Guilyardi, E.: Late-twentieth-century emergence of the El Niño propagation asymmetry and future projections, Nature, 504, 126–130, doi:10.1038/nature12683, 2013.

Solow, A. R., Adams, R. F., Bryant, K. J., Legler, D. M., O'Brien, J. J., McCarl, B. A., Nayda, W., and Weiher, R.: The Value of improved ENSO prediction to U.S. agriculture, Climatic Change, 39, 47–60, 1998.

Sterl, A., Oldenborgh, G. J., Hazeleger, W., and Burgers, G.: On the robustness of ENSO teleconnections, Clim. Dynam., 29, 469–485, 2007.

Timm, O.: Nonstationary ENSO-precipitation teleconnection over the equatorial Indian Ocean documented in a coral from the Chagos Archipelago, Geophys. Res. Lett., 32, doi:10.1029/2004GL021738, 2005.

Tziperman, E., Zebiak, S. E., and Cane, M. A.: Mechanisms of seasonal – ENSO interaction, J. Atmos. Sci., 54, 61–71, 1997.

van Oldenborgh, G. J. and Burgers, G.: Searching for decadal variations in ENSO precipitation teleconnections, Geophys. Res. Lett., 32, doi:10.1029/2005GL023110, 2005.

Vecchi, G. A. and Wittenberg, A. T.: El Niño and our future climate: where do we stand?, Wiley Interdisciplinary Reviews: Climate Change, 1, 260–270, 2010.

von Storch, H., Zorita, E., and González-Rouco, F.: Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation, Int. J. Earth Sci., 98, 67–82, 2009.

Wang, C., Deser, C., Yu, J.-Y., DiNezio, P., and Clement, A.: El Niño and Southern Oscillation (ENSO): A Review, Spring Science Publisher, Berlin, 2012.

Watanabe, M., Kug, J.-S., Jin, F.-F., Collins, M., Ohba, M., Wittenberg, A. T.: Uncertainty in the ENSO amplitude change from the past to the future, Geophys. Res. Lett., 39, doi:10.1029/2012GL053305, 2012.

Wilson, R., Cook, E., D'Arrigo, R., Riedwyl, N., Evans, M. N., Tudhope, A., and Allan, R.: Reconstructing ENSO: the influence of method, proxy data, climate forcing and teleconnections, J. Quaternary Sci., 25, 62–78, 2010.

Wittenberg, A. T.: Are historical records sufficient to constrain ENSO simulations?, Geophys. Res. Lett., 36, doi:10.1029/2009GL038710, 2009.

Wittenberg, A. T.: Variation of ENSO Teleconnections, 2012 AGU Fall Meeting, San Francisco, California., Abstract OS52B-07, 7 December 2012.

Wittenberg, A. T., Rosati, A., Lau, N.-C., and Ploshay, J. J.: GFDL's CM2 global climate models. Part III: Tropical pacific climate and ENSO, J. Climate, 19, doi:10.1175/JCLI3631.1, 2006.

Xiao, M., Zhang, Q., and Singh, V. P.: Influences of ENSO, NAO, IOD and PDO on seasonal precipitation regimes in the Yangtze River basin, China, Int. J. Climatol., doi:10.1002/joc.4228, 2014.

Yeh, S.-W., Kug, J.-S., and An, S.-I.: Recent progress on two types of El Niño: observations, dynamics, and future changes, Asia-Pac. J. Atmos. Sci., 50, 69–81, 2014.

Zorita, E., González-Rouco, F., and Legutke, S.: Testing the Mann et al. (1998) Approach to Paleoclimate Reconstructions in the Context of a 1000-Yr Control Simulation with the ECHO-G Coupled Climate Model, ~~J. Climate, 16, 1378–1390, 2003.~~ Science, 339, 67–70, 2013.

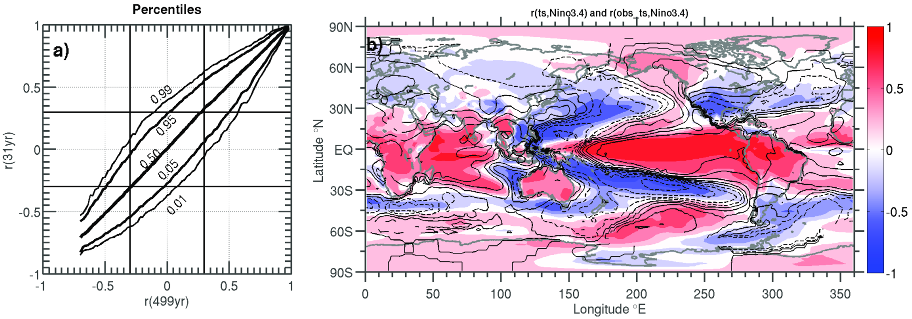

Figure 1. (a) The percentiles of correlations found in 31 year segments between the model (see Sect. 2) surface temperature (TS) at each grid point and the model calculated Nino 3.4 index (y axis), plotted against the corresponding correlations for the whole 499 years of data (x axis). The lines are the 1st, 5th, 50th, 95th, and 99th percentiles, with the lowest lines indicating the lowest percentiles (i.e. the bottom line is the 1st percentile). **(b)** The shading is the correlation between of the entire 499 years of TS at each grid point and the model calculated Nino 3.4 index~~correlation coefficients~~, both calculated from the GFDL CM2.1 data, also described in Sect. 2. The black contour lines are the correlation coefficients (spacing of 0.2) of observed surface land-sea temperatures to its corresponding Nino 3.4. Soild lines are positive values, while dashed lines are negative. These observations were calculated using the last 50 years of annual mean GISTEMP_ersst observational data (GISTEMP-Team, 2015). Dataset is described by Hansen et al. (2010).

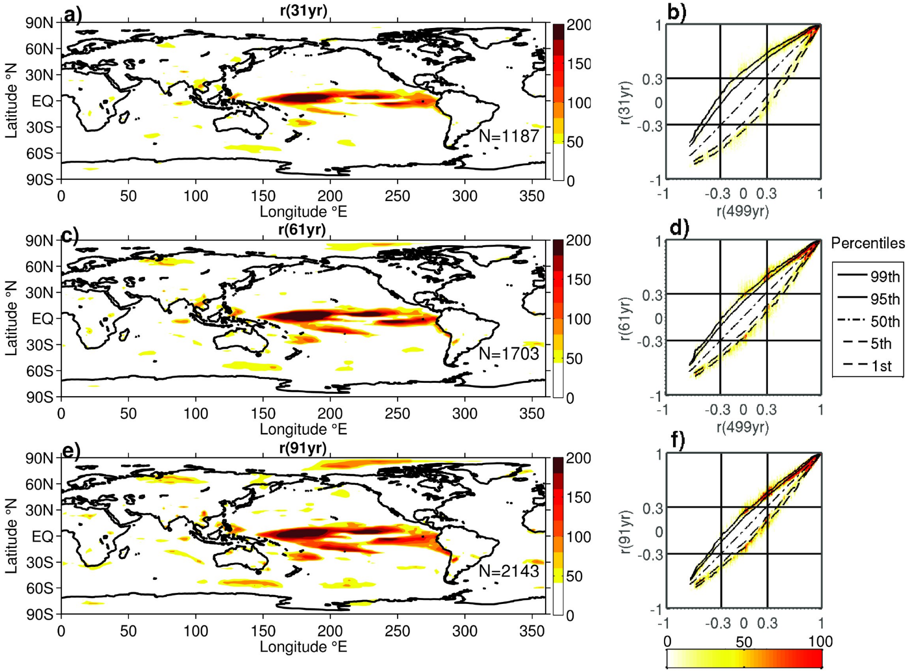

Figure 2. Panels **(a)**, **(c)** and **(e)** show the number of non-stationary ~~years~~ windows (coloured shading) for each grid point over the entire dataset for 31, 61 and 91 year windows, respectively. ~~The yellow to red values are defined as non-stationary according to Sect. 3, and have~~ This shading has been adjusted for the slightly different lengths of data available for the different calibration window length. The number of non-stationary grid points (using 499 years of data) for any window is shown in bottom right corner of each panel as N. Panels **(b)**, **(d)** and **(f)** shows the percentiles of correlations between global TS and Nino 3.4 in 31, 61 and 91 year windows respectively (y axis), verses the corresponding correlations for the whole 499 years of data (x axis). This plot is very similar to Fig. 1a, but with the underlying coloured shading representing the y axis positions of non-stationary ~~years~~ windows in the plot (according to definition of non-stationarity, see Sect. 3). A deeper red indicates a higher density of points, as many points can occupy the same correlation values.

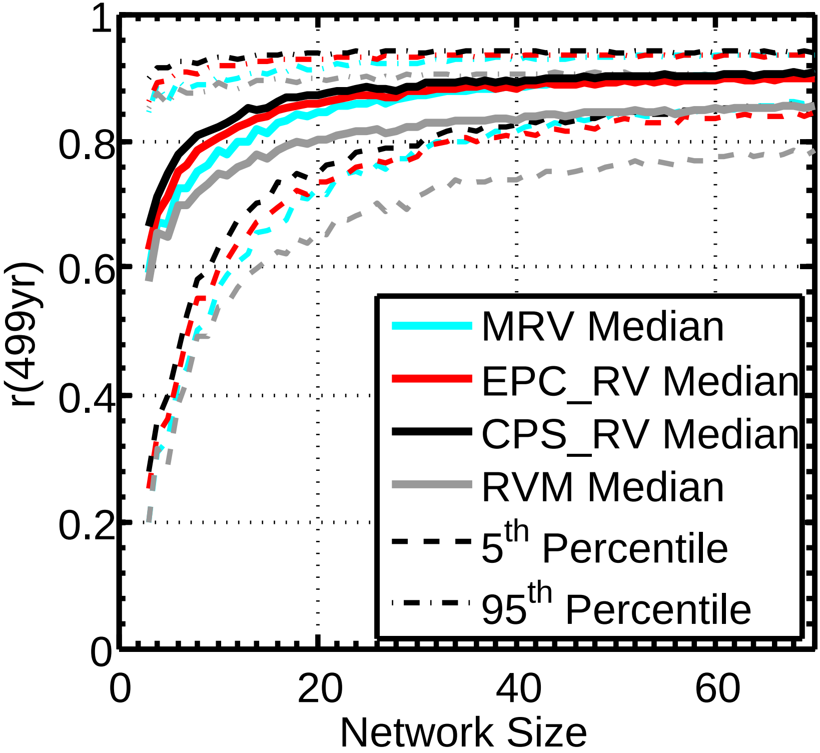

Figure 3. The 5th (lower dashed), 50th (thick) and 95th (upper dotted and dashed) percentiles of correlation coefficients calculated between the pseudo-reconstructions of running variance and ENSO running variance (y axis) plotted against the proxy network size (x axis). The percentiles are calculated across the 1000 iterations of randomly selected groups of source proxies. These reconstructions are from the first series of experiments which involve using the entire 499 years of data, for more information see Sect. 3.

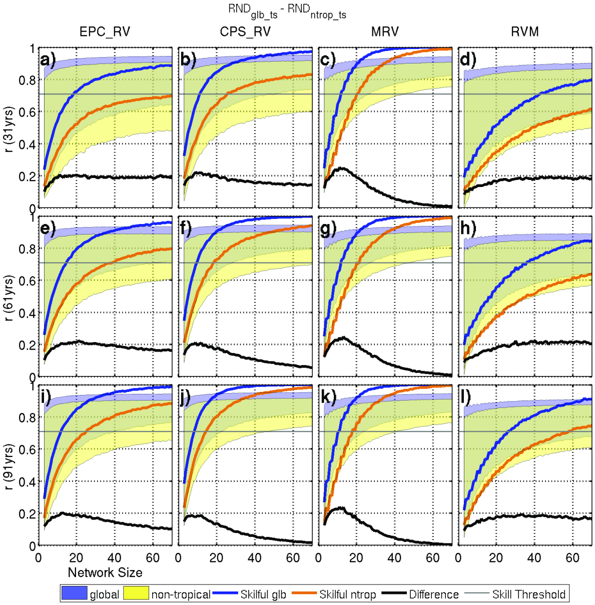

Figure 4. A comparison of reconstruction skill of the global $\text{RND}_{\text{glb\_ts}}$ (blue) and the non-tropical $\text{RND}_{\text{ntrop\_ts}}$ (yellow) experiments. Correlation coefficients are calculated between the ~~reconstruction's~~ reconstructed running variance and ENSO running variance (y axis), plotted against the proxy network size (x axis). The coloured regions show the range of these coefficients, from the 5th to the 95th percentile, with overlapping regions shown by the yellow-green colouring. The thick blue and orange lines show the proportion of skilful reconstructions for the $\text{RND}_{\text{glb\_ts}}$ and $\text{RND}_{\text{ntrop\_ts}}$ experiments respectively. Skilful reconstructions are defined as explaining greater than 50 % of ~~explained~~ variance (grey line). The black line is the difference in skill between the $\text{RND}_{\text{glb\_ts}}$ (blue line) and $\text{RND}_{\text{ntrop\_ts}}$ (orange line) experiments. Each row corresponds to different calibration window lengths, titled on the y axis. Each column represents different reconstruction methods, titled at the top of the columns.

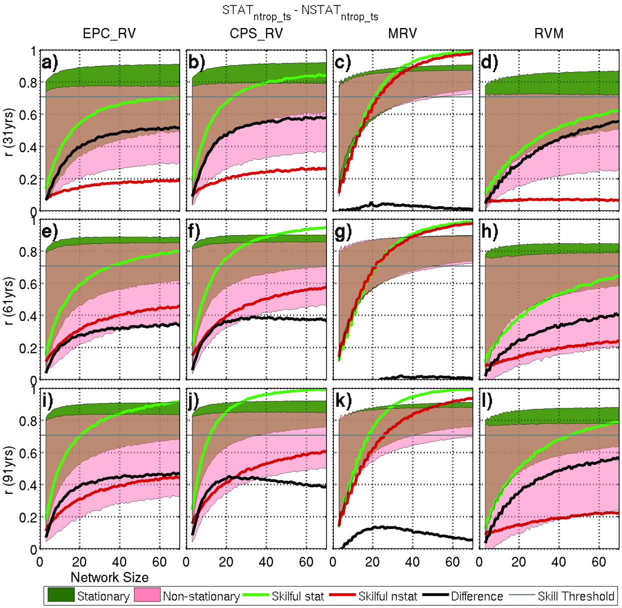

Figure 5. A comparison of reconstruction skill of the "stationary" STAT$_{\text{ntrop\_ts}}$ (green) and non-stationary NSTAT$_{\text{ntrop}}$ (pink) experiments. Correlation coefficients are calculated between the ~~reconstruction's~~ reconstructed running variance and ENSO running variance (y axis), plotted against the proxy network size (x axis). The coloured regions show the range, from the 5th to the 95th percentile, with overlapping regions shown by the brownish colouring. The thick green and red lines show the proportion of skilful reconstructions for the STAT$_{\text{ntrop}}$ and NSTAT$_{\text{ntrop}}$ experiments respectively. Skilful reconstructions are defined as explaining greater than 50 % of explained variance (grey line). The black line is the difference in skill between the STAT$_{\text{ntrop}}$ (green line) and NSTAT$_{\text{ntrop}}$ (red line) experiments. Each row corresponds to different calibration window lengths, titled on the y axis. Each column represents different reconstruction methods, titled at the top of the columns.

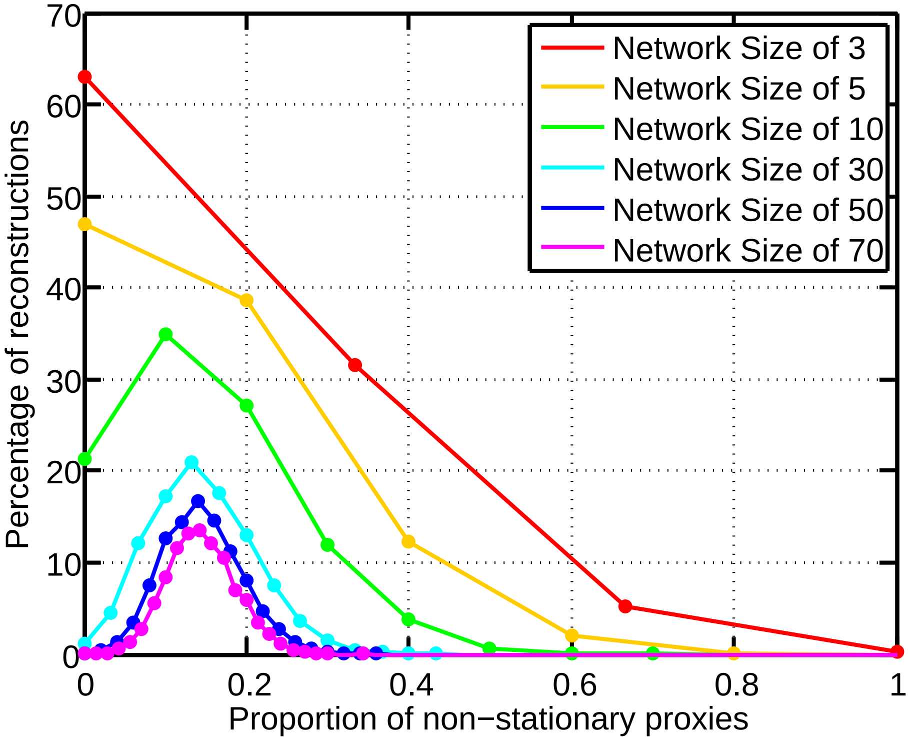

Figure 6. This plot shows the percentage of TS based reconstructions (y axis) with certain proportions of non-stationary proxies (x axis) for the $\mathrm{RND_{glb\_ts}}$ experiment. Each of the ten 31 year calibration windows has been included in these calculations, so that the proportions of non-stationarities for 10 000 reconstructions are shown (50 % being 5000 reconstructions). Different lines are for different proxy network sizes (see inset legend), and this determines what values of proportion can be taken~~as~~, hence larger groups have a wider range of possible non-stationarity proportions than smaller groups. The coloured circles of any proportions with 0 % of reconstructions have not been shown.

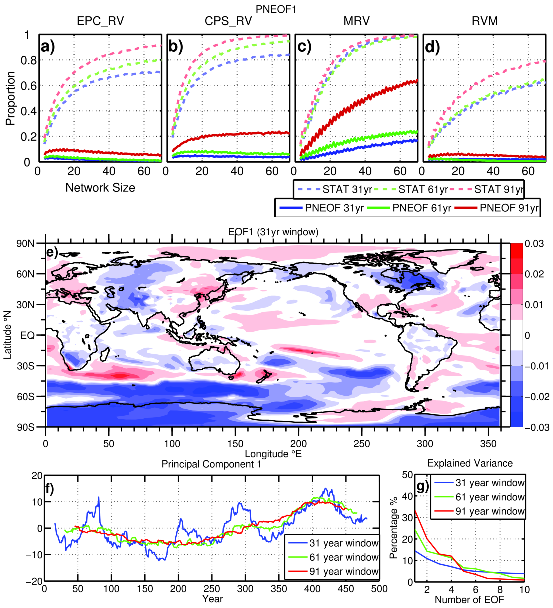

Figure 7. (a–d) The reconstructions from the PNEOF1 (solid) and $\mathrm{STAT}_{\mathrm{ntrop\_ts}}$ (dashed) experiments using different reconstruction methods. The proportion of skilful reconstructions (out of the 10 000) is shown for calibration windows of 31 (blue), 61 (green) and 91 years length (red), plotted against proxy network size (x axis). Skilful is defined as the reconstruction explaining greater than 50 % of the variance of the model ENSO reconstruction. **(e)** The spatial map of the leading Empirical Orthogonal Function (EOF) of running correlations calculated between TS at each grid point and ENSO (with a window length of 31 years). The spatial structure of this EOF is quantitatively similar to the first EOF with running correlation window lengths of 61 (spatial $r = 0.86$), and 91 (spatial $r = 0.84$) years. **(f)** The leading principal components of the leading EOF with running correlation window lengths of 31 (blue), 61 (green) and 91 (red) years. The year values correspond to the centre of the windows. **(g)** The variance explained by the first 10 EOFs, for the three different window sizes (see inset legend).

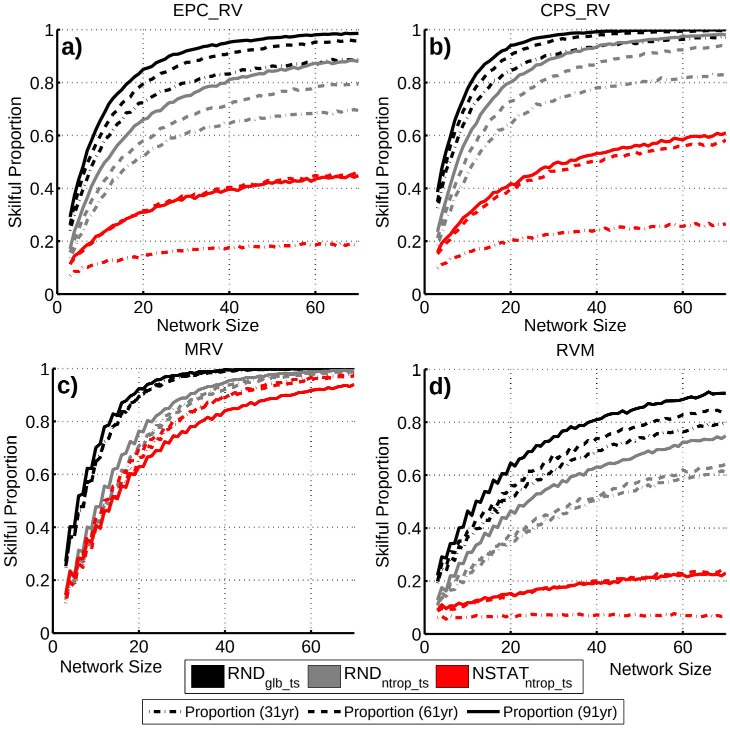
EPC_RV
a)
Skilful Proportion
0
0.2
0.4
0.6
0.8
1
0
20
40
60
Network Size
CPS_RV
b)
MRV
c)
RVM
d)
RND glb_ts
RND ntrop_ts
NSTAT ntrop_ts
Proportion (31yr)
Proportion (61yr)
Proportion (91yr)

Figure 8. Reconstruction skill for different experiments (~~each row~~as indicated by colour) and reconstruction methods (each ~~column~~panel) using different calibration window lengths ~~. Correlation coefficients are calculated between the reconstruction's running variance and ENSO running variance (y axis), plotted against the proxy network size (x axis). This is done for calibration windows of 31 (blue hatching), 61 (green hatching) and 91length~~ (~~red hatching~~see inset legend). The ~~hatching shows the range from the 5th percentile to the 95th percentile of the correlation coefficients. The thick blue, green, and red~~ lines show the proportion of skilful reconstructions ~~for the three calibration windows~~ with skilful being ~~31, 61 and 91length respectively. Skilful reconstructions are~~ defined as explaining greater than 50 % of the explained variance ~~(grey line)~~of the Niño 3.4 running variance.

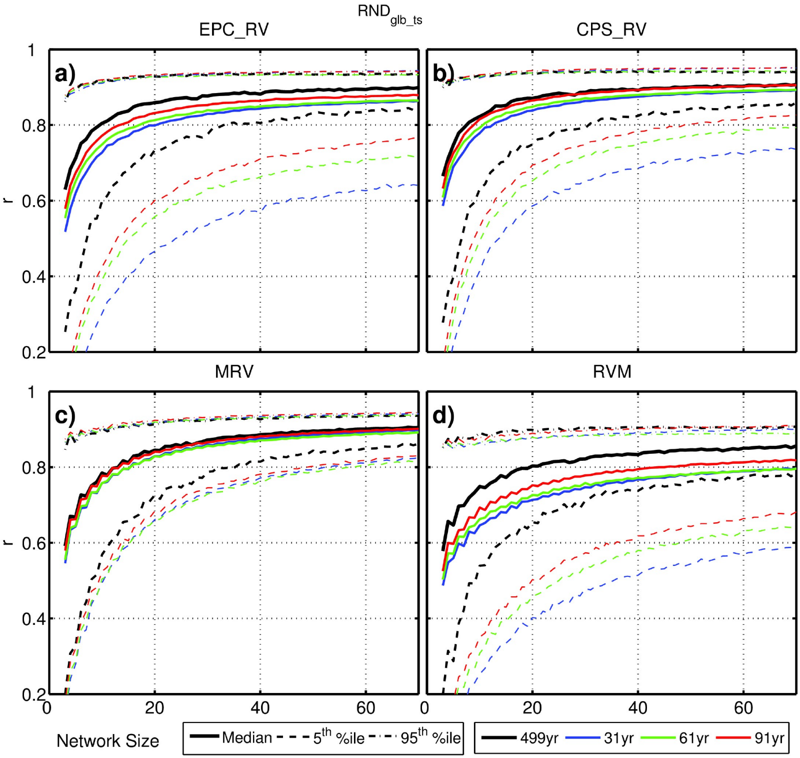

Figure 9. A comparison of all the RND_{glb_ts} reconstructions using 499 years of data (black) and when using limited calibration windows of 31 (blue), 61 (green), and 91 years (red). The 5th (dashed), 50th (solid line) and 95th (dot-dashed) percentiles of correlation coefficients are shown for each of the window lengths and for reconstructions using the 499 years of data. Correlation coefficients are calculated between the ~~reconstruction's~~ reconstructed running variance and ENSO running variance (y axis), plotted against the proxy network size (x axis). Panels **(a–d)** show the comparison for the four reconstruction methods discussed in Sect. 3.3.

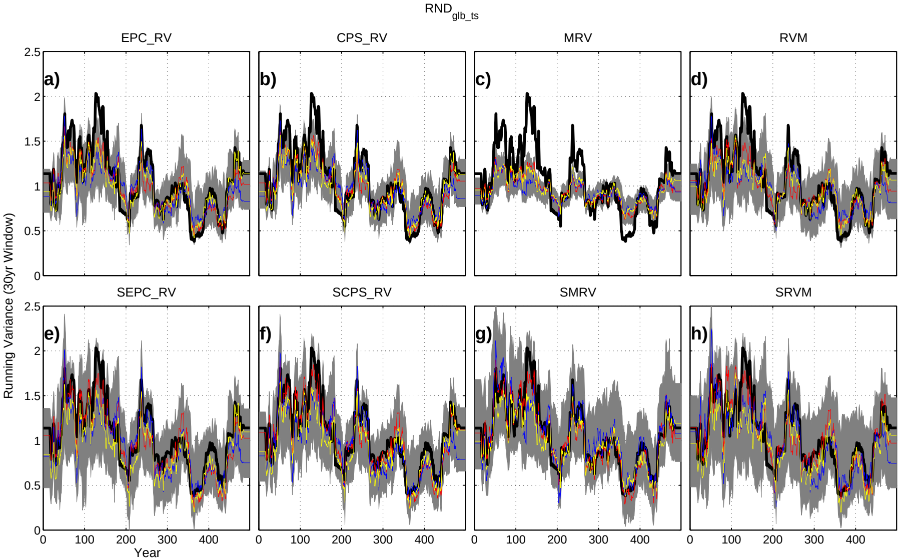
RND$_{glb\_ts}$
EPC_RV
CPS_RV
MRV
RVM
SEPC_RV
SCPS_RV
SMRV
SRVM
a)
b)
c)
d)
e)
f)
g)
h)
Running Variance (30yr Window)
Year
0
0.5
1
1.5
2
2.5
0
100
200
300
400

Figure 10. ~~Panels **(a)**, **(c)**, and **(e)** show Nino~~ The reconstructed Niño 3.4 running variance ~~reconstructions~~ range of various methods (~~green~~grey shading) plotted ~~alongside the standard deviations of~~ with the ~~running correlations (30window) of its constituent proxies to~~ model ~~Nino~~ Niño 3.4 ~~(blue)~~running variance. The first row of panels are the unscaled reconstruction methods, while the second row has been scaled with a linear regression. Reconstruction methods are indicated in the title above the panel, and the scaled variants are named beginning with an 'S'. The year of the reconstruction is shown on the x axis of these panels. ~~One~~ Running variances are calculated using 30 year moving windows. Some example ~~of each reconstruction method is shown~~ reconstructions (~~indicated in the title)~~red, yellow and ~~the correlation between the~~ blue ~~and green~~ lines) are shown~~in the bottom right corner of each panel. The red circles suggest when there is a clear negative correlation between the time series, discussed in Sect. 5. Panels **(b)**, **(d)** and **(f)** show~~ with the colours corresponding ~~scatterplots~~ to a specific network of pseudoproxies. These reconstructions are from the ~~two time series in panels **(a)**, **(d)**, and **(e)** respectively, such that the standard deviation lies on the x axis~~RND_{glb_ts} experiment, ~~and the Nino 3.4 variance reconstruction is on the y axis. A least-squares-fit line is provided~~ with ~~each scatterplot (blue line)~~a 91 year calibration window.