Quantifying paleo-reconstruction skill of the Southern Annular Mode in a model framework

Willem Huiskamp[1,2] and Shayne McGregor[3]

[1]RD1- Earth System Analysis, Potsdam-Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Brandenburg, Germany
[2]Climate Change Research Centre, UNSW Sydney, Sydney, NSW, Australia
[3]School of Earth Atmosphere and Environment, Monash University, Melbourne, Victoria, Australia

Correspondence: Willem Huiskamp (huiskamp@pik-potsdam.de)

Abstract. Past attempts to reconstruct the Southern Annular Mode (SAM) using paleo archives have resulted in records which can differ significantly from one another prior to the window over which the proxies are calibrated. This study attempts to quantify not only the skill with which we may expect to reconstruct the SAM, but also assess the contribution of regional bias in proxy selection and the impact of non-stationary proxy-SAM teleconnections on a resulting reconstruction. This is achieved using a pseudo-proxy framework with output from the GFDL CM2.1 global climate model. Reconstructions derived from precipitation fields perform better, with 89% of reconstructions calibrated over a 61 year window able to reproduce at least 50% of inter-annual variance in the SAM, as opposed to just 25% for surface air temperature (SAT) derived reconstructions. Non-stationarity of proxy-SAM teleconnections, as defined here, plays a small role in reconstructions, however the range in reconstruction skill is not negligible. Reconstructions are most likely to be skilful when proxies are sourced from a geographically broad region, with a network size of at least 70 proxies.

Copyright statement. ©Authors 2020. This work is distributed under the Creative Commons Attribution 4.0 License.

1 Introduction

The Southern Annular Mode (SAM) is the leading mode of atmospheric variability in the Southern Hemisphere which describes the intensity and latitudinal location of the subtropical westerly jet. Positive and negative phases of the SAM have been linked to changes in surface air temperature (SAT) and precipitation in Australia (Hendon et al., 2007), New Zealand (Gallant et al., 2013) as well as South America and Africa (Gillet et al., 2006; Silvestri and Vera, 2009). For example, positive phases of the SAM over the period 1979-2005 are typically associated with cool annual temperature anomalies over the Antarctic continent (Thompson and Solomon, 2002; Kwok and Comiso, 2002; Gillet et al., 2006) and warm anomalies over the Antarctic Peninsula, southern South America and southern New Zealand (Kwok and Comiso, 2002; Silvestri and Vera, 2009). Precipitation changes typically found during a positive SAM phase include negative annual precipitation anomalies over southern South America, New Zealand and Tasmania and positive precipitation anomalies over Australia and South Africa (Silvestri and Vera, 2009).

Over the last five decades the SAM has shown a trend towards more positive values, consistent with a poleward intensification of the surface westerly winds that has been largely attributed to anthropogenic forcing, such as stratospheric ozone depletion and the increase in atmospheric CO_2 (Son et al., 2008; Lee and Feldstein, 2013; Previdi and Polvani, 2014). In addition, both high frequency (3-4 months) and low frequency (16 years) variability has been observed in the SAM as derived from reanalysis experiments (Raphael and Holland, 2006). It is important to place these observed multi-decadal trends over the last five decades into a long-term context in order to understand the contributions of forced and natural variability. These relative contributions are important for understanding projected future changes given the impact of the SAM not only on regional weather patterns, but also large scale ocean circulation and heat uptake (Russell et al., 2006; Marini et al., 2011; Liu et al., 2018), and even the marine carbon cycle (Lovenduski et al., 2007; Lenton and Matear, 2007; Le Quéré et al., 2007; Huiskamp and Meissner, 2012; Hauck et al., 2013; Huiskamp et al., 2015; Keppler and Landschützer, 2019).

Instrumental reconstructions of the SAM extend as far back as 1865 (Jones et al., 2009), but significant uncertainty exists prior to the mid 20^{th} Century due to fewer observations and the methods used to compensate for this (i.e, estimates based on atmospheric conservation of mass, such as Jones et al. (2009)). Direct measurements, meanwhile, only extend as far back as 1958 (Marshall, 2003). Thus, if we wish to extend our understanding of SAM trends and variability back beyond the instrumental record, reconstructions derived from paleo archives are required.

1.1 Paleo reconstructions of SAM variance

Paleo-reconstructions are generated by examining changes preserved in natural environmental archives (biological, chemical and physical records) that are sensitive to climatic impacts of the mode of variability being reconstructed. In the case of the SAM, this has traditionally been achieved by finding proxies that are sensitive to precipitation or surface air temperature changes associated with the two different phases of the SAM. Proxies recording changes in temperature and precipitation include tree rings, ice cores and terrestrial sediment cores, although the latter are less favoured due to chronological uncertainties and a typically lower temporal resolution. Growth of trees, and therefore ring width and/or density, can be sensitive to both temperature and precipitation depending on the tree type and its location, while ice cores can provide accumulation rates, $\delta^{18}O$ and δD (e.g. Steig et al. (2005)) - which record air temperature and precipitation (accumulation).

1.1.1 Reconstructions and their potential issues

The relationship to the SAM is typically initially established by correlating changes in these proxy records with a SAM index developed from instrumental or reanalysis data over a period spanning several decades (Villalba et al. (2012) and references therein; Abram et al. (2014)). The individual proxy records are then combined into a single index using a reconstruction method such as a regression approach (Zhang et al., 2010; Villalba et al., 2012) or weighted Composite Plus Scaling (CPS; Abram et al., 2014; Dätwyler et al., 2018).

There are, however, two fundamental assumptions being made when we reconstruct past climate in this way. Firstly, we assume that a hemisphere-wide mode of variability can be accurately reconstructed using records from a geographically limited sample space. As there is relatively little land in the Southern Hemisphere, particularly in the latitude of the westerlies, SAM reconstructions often rely disproportionately on records from narrow longitude bands. Sites are primarily in South America, Australia and New Zealand (Villalba et al., 2012), and Antarctica (Zhang et al., 2010; Abram et al., 2014; Dätwyler et al., 2018), with Antarctica being the only location able to provide samples with good longitudinal coverage. Abram et al. (2014) suggest their regional Drake Passage sector paleo SAM reconstruction is representative of the hemispheric mean signal by extracting a sea level pressure-derived (SLP) SAM index from a suite of 8 global climate models and comparing it with a secondary SAM index that is derived from the same SLP field, but restricted to the Drake Passage sector. They find that the regional expression of the SAM in these models closely resembles the hemispheric expression over 1000 years, a conclusion supported by the regional SAM records of Visbeck (2009). Dätwyler et al. (2018) on the other hand find non-trivial differences between their hemisphere-wide SAM reconstruction and that of Abram et al. (2014), implying that an annual-mean SAM reconstructed from paleo proxies is not well approximated by sampling from a limited region.

Secondly, when we correlate a proxy to the modern SAM over a calibration window of several decades, we make the assumption that this relationship remains the same through time. This is commonly referred to as proxy stationarity. Gallant et al. (2013) investigated SAT/precipitation non-stationarity using instrumental data spanning the period 1900-2009 and reported that 21-37% of Australian precipitation records showed non-stationary teleconnections to the El Nino-Southern Oscilation (ENSO) and the SAM. Silvestri and Vera (2009) performed a similar study with observed precipitation and surface air temperature records from Australia and South America spanning the 1960-70's and 1980-90's spring months. They found that significant positive correlations of the SAM with SAT in the Australia/ New Zealand region in the earlier decades can become insignificant or even negative in the more recent decades. Dätwyler et al. (2018) built on this by adding a stationarity criteria to their proxies for reconstructing the SAM, but at a cost of calibrating their proxies with a longer, but less reliable record (Jones et al., 2009). The resulting reconstructions showed a more stable teleconnection through time, but were not necessarily more skilful (as measured by validation statistics). Finally, when considering multi-decadal calibration periods, stochastic noise or other climate signals (e.g. ENSO) can modulate the correlation strength between, for example, South American precipitation and the SAM without the precipitation record being classified as non-stationary (Yun and Timmermann, 2018).

This study aims to quantify the uncertainties raised by the aforementioned assumptions within a modelling framework, similar to Batehup et al. (2015), and seeks to address the following questions: 1) What impact does proxy network size and calibration window size have on the skill of a resulting reconstruction? 2) How does the geographical distribution of the proxies affect reconstruction skill? 3) Are any regions in our model framework prone to producing non-stationary proxies and what could be modulating the SAM-proxy teleconnection? The use of climate models to assess the skill of paleo reconstructions provides an opportunity to investigate a 'perfect' time-series of the climate index we wish to reconstruct and the ability to reconstruct this index with fields from a model, which act as pseudo paleo-proxies. These 'perfect' proxies are free from non-climatic noise that may degenerate a teleconnection signal in a real proxy, isolating instead changes in teleconnection strength due to underlying variability in the climate. This is in contrast to 'real world' proxies which are also prone to other influences inherent with the physical/chemical/biological nature of the proxy itself. It is often assumed that these effects will be minimised by sampling proxies from a range of regions as local factors would not be expected to be correlated amongst differing locations. Additionally, model data allows us to assess multi-decadal to centennial changes in proxy-SAM teleconnection and how calibration over certain windows in time affects the skilfulness of a SAM reconstruction.

2 Methods

2.1 Model and data

The data used in this study are a 500-year pre-industrial control simulation of the Geophysical Fluid Dynamics Laboratory's Coupled Model 2.1 (CM2.1 hereafter), with all boundary conditions set to CE 1860 levels. This assures that any changes in the model SAM are due to internal variability in the climate system only. The model is a fully-coupled global climate model with ocean (OM3.1), atmosphere (AM2.1), land (LM2.1) and sea ice components (SIS). The ocean model has a resolution of 1°x 1° which increases equatorward of 30° to a meridional resolution of 1/3° at the equator (Griffies et al., 2005). The atmospheric and land surface models have a resolution of 2° latitudes by 2.5° longitude and the AM2.1 has 24 vertical levels (Delworth et al., 2006).

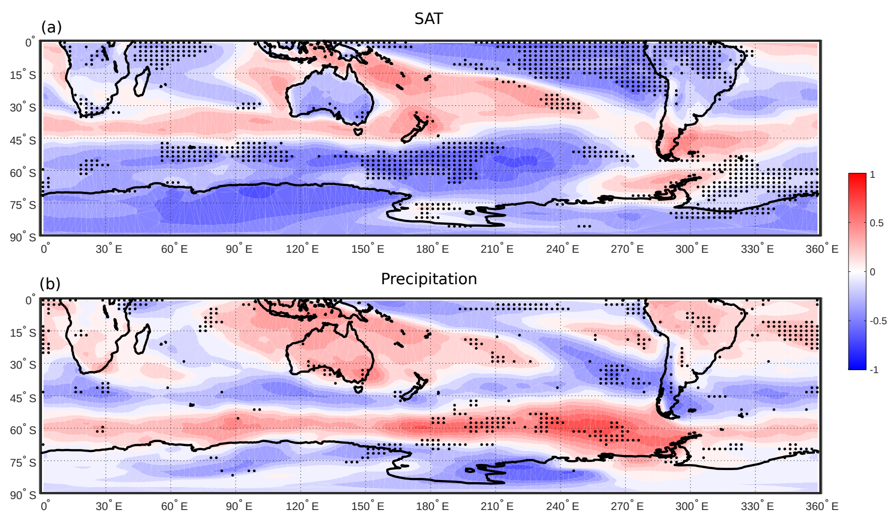

Figure 1. Correlations of annual-mean (Jan-Dec) SAT (a) and precipitation (b) from the GFDL CM2.1 model with the model-derived SAM, calculated over 500 years. Black dots show where the correlation of the ERA-Interim reanalysis product with the Marshall SAM index, calculated over a 36 year period from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

CM2.1 is selected due to its good representation of the SAM compared to similar models from the CMIP5 and CMIP6 archives (Bracegirdle et al., 2020) while Karpechko et al. (2009) find performance to be favourable when compared to ERA-40 data. The spatial structure of the SAM is well simulated, accurately capturing the centre of action over the Pacific, while being slightly too zonally symmetric on the eastern half of the Southern Hemisphere (Raphael and Holland, 2006). Importantly for our purposes, CM2.1 accurately simulates the latitude at which the SAM transitions from its positive to its negative phase (as expressed via regression onto 850hPa winds) over South America, which many models of a similar age and computational complexity fail to achieve (Raphael and Holland (2006), their Figure 4b). The amplitude of the model SAM index is comparable with observations (Raphael and Holland, 2006), although its variability is larger than observed (Karpechko et al., 2009). As previously noted, we should be cautious directly comparing observations (in this instance, ERA-Interim (Dee et al., 2011) data, correlated with the Marshall SAM index (Marshall, 2003) over the 36 year period from 1979-2014) spanning a brief time period with our model data which spans 500 years. To address this, we calculate a 36 year running correlation between the model SAM and our SAT and precipitation fields and identify if the correlations derived from observations fall within the model range (Figure 1). The SAM index in the model is calculated according to the method of Gallant et al. (2013) as the difference in normalized, zonally averaged sea level pressure anomalies between 40°S and 60°S. Aside from a region in equatorial South America in the SAT field, the agreement is good, with 87% of SAT and 95% of precipitation grid cells on land showing agreement with observations.

Paleo-proxies are not uniformly sensitive to one season or variable, depending on the region from which they are sourced. For example, a tree ring record constructed by Cullen and Grierson (2009) from south-west Western Australia is shown to be particularly sensitive to austral autumn-winter rainfall. Alternatively, South American tree ring records compiled by Villalba et al. (2012) show sensitivity to summer-autumn precipitation, while New Zealand records appear to be most responsive to summer temperature. In addition, while the proxies are sensitive to SAT or precipitation during one season, the SAM's strongest influence on these variables may be during a different season entirely. For example, while the South American tree rings of Villalba et al. (2012) are sensitive to summer-autumn precipitation, the SAM signal is most clearly seen in late spring and winter precipitation in south-eastern South America (Silvestri and Vera, 2009). For the sake of simplicity, we employ annual mean (Jan-Dec) fields for sea level pressure, surface air temperature and precipitation and focus instead on the impact of network size and calibration window length rather than seasonal effects.

2.2 Calculation of non-stationarity

A proxy is typically considered non-stationary if its teleconnection to the SAM is changed by some dynamical process rather than stochastic variability (localised weather) such that the signal it records is no longer representing changes in the SAM. Here, the SAM teleconnection is modelled via a running correlation between the proxy and the SAM index over a window of 31, 61 or 91 years. We define non-stationary teleconnections following the method of Gallant et al. (2013) and Batehup et al. (2015), such that a proxy is considered non-stationary when the variability in its running correlation with the SAM exceeds what would be expected if the proxy were only influenced by local random noise.

Following from this, a Monte Carlo approach (van Oldenborgh and Burgers, 2005; Sterl et al., 2007; Gallant et al., 2013) is used to create stochastic simulations of SAT and precipitation at each grid point in the model. These stochastic simulations are created to have the same statistical properties as the original SAT and precipitation data from the CM2.1 simulation. To determine the range of variability expected due to the stochastic processes mentioned previously, one thousand of these time series are created at each grid point according to the following equation from Gallant et al. (2013):

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

$v(t)$ is the stochastic SAT or precipitation time series. a_0 and a_1 are regression coefficients representing the stationary teleconnection strength between SAT or precipitation and the SAM index $(c(t))$ while the remaining terms represent the noise added to the time-series. A red noise, $[\eta_v(t) + \beta\eta_v(t-1)]$, is added and weighted by the standard deviation σ_v of the local SAT or precipitation time-series as well as the proportion of the variance not related to the regression ($\sqrt{1-r^2}$), where r is the correlation between the SAT/precipitation time-series and the SAM index. The red noise is a combination of random Gaussian noise ($\eta_v(t)$) and autocorrelation (β) of the SAT or precipitation time-series at a lag of one year multiplied by the Gaussian noise ($\beta\eta_v(t-1)$).

The stochastic simulations of SAT and precipitation are used to create a 95% confidence interval for each grid point of all possible running correlations a time-series could have and still be considered to have a stationary teleconnection with the SAM. Therefore, if the time-series from our model proxy has a running correlation that falls outside the confidence interval, we consider that proxy to be non-stationary with the SAM in that temporal window, as it is unlikely to be affected by stochastic processes alone. It should be noted that as a 95% confidence interval is used, non-stationarity will be falsely identified 5% of the time, hence we define a grid point as non-stationary only if the running correlation falls out of the confidence interval more than 10% of the time, or 50 of the 500 years; more than double the 5% we might expect by chance alone. As correlations are bounded between ± 1, the running correlations are converted to Fisher Z-scores:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \tag{2}$$

where r is the running correlation value.

2.3 Generation of pseudoproxies

SAT and precipitation fields from the model are used to represent climate proxies in the model, as discussed in Section 2.1. Rather than being inferred via changes in tree ring growth, these proxies are direct measures of these variables and therefore free of non-climatic noise (von Storch et al., 2009). We do not add noise to increase the realism of these proxies, rather we assess reconstruction skill and non-stationarity in a ‘best-case-scenario’ where we assume the proxy is a perfect analogue for the climate variable it is deemed to represent (SAT or precipitation), similar to the experiments of Dätwyler et al. (2020).

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

Ideally, proxies would be calibrated with the SAM over the full length of the time-series - 500 years, however as previously noted, real world proxies are calibrated over shorter windows of several decades. For each gridpoint in the model, the time-series is split into 10 windows of either 31, 61 or 91 years in length, whose midpoints are evenly spaced throughout the 500 years, regardless of overlap or space between them. The number of proxies meeting our aforementioned criteria in each region/windowsize can be found in Table 1. This approach allows for the possibility that a proxy may have a strong correlation with the SAM over the calibration window, but which may be insignificant or even reversed over other windows, or indeed the full 500 years. These window sizes are chosen to assess the effect differing window lengths have on the resulting skill of the reconstruction. For example, the use of a 61 year calibration window, as opposed to a 31 year one, may decrease the effect of decadal climate variability and its modulation of the pseudoproxy-SAM teleconnection.

Reconstructions are computed with a network size of between 2 and 70 proxies, a range typical of past reconstructions with strict selection criteria (eg. Abram et al. (2014) and Dätwyler et al. (2018)). This is done to quantify the dependence of reconstruction skill on network size. 1000 networks are generated for each of the 10 calibration windows for each network size. Each site in each network is randomly selected and unique, while the same site may be included in more than one network.

To reconstruct the proxy networks into a single proxy-SAM index, we use the weighted composite plus scale (CPS) method (Esper et al., 2005; Hegerl et al., 2007), similar to that used by Abram et al. (2014). As the scope of this study does not include the effect of different reconstruction methods on the skill of the reconstructed index, CPS is used as it is commonly employed in paleo-reconstructions (PAGES 2k Consortium, 2013; Abram et al., 2014; Batehup et al., 2015; Dätwyler et al., 2018). Using this method, proxies are normalised to have a 0 mean, unit standard deviation and then weighted according to their correlation to the model SAM over the calibration window, before being summed into a single time-series. To quantify the skill of the pseudoproxy reconstructions, Pearson correlation coefficients are calculated between each normalised SAT/precipitation-derived SAM index and the sea level pressure-derived SAM index over the full 500 years of data. We define a skilful reconstruction as one that is able to reproduce at least 50% of the model SAM variability (i.e., $r^2 \geq 0.5$ or $r \geq\sim 0.71$).

Running correlations of SAT/precip and the SAM are correlated with the model Nino3.4 (n3.4) index to investigate the role ENSO may play in modulating the pseudoproxy-SAM teleconnection. The model n3.4 index is calculated as the sea surface temperature anomaly in the region bounded by 5°N to 5°S and 170°W to 120°W.

Each grid SAT and precipitation grid cell is correlated with the SAM over a 31, 61 and 91 year running window, while the n3.4 index is band-pass-filtered using the same window size to remove high-frequency variability. The two time-series are then correlated over their common interval (500 years - window size/2) with significance calculated using a reduced degrees of freedom method (Davis, 1976). An additional set of SAM reconstructions are calculated which exclude any proxy whose SAM-proxy running correlation is found to have a significant ($p < 0.05$) correlation to the filtered n3.4 index over the relevant calibration window.

Region	Window size	Number of sites	
		SAT	Precip
S. Hemisphere	31yrs	842 - 1740	549 - 935
	61yrs	640 - 1568	429 - 709
	91yrs	838 - 1535	326 - 660
Antarctica	31yrs	557 - 1346	264 - 563
	61yrs	454 - 1253	191 - 403
	91yrs	705 - 1254	211 - 396
Aus/NZ	31yrs	48 - 152	60 - 166
	61yrs	41 - 130	46 - 165
	91yrs	31 - 132	33 - 158
S. America	31yrs	54 - 244	62 - 195
	61yrs	44 - 227	46 - 156
	91yrs	30 - 207	39 - 154

Table 1. The range of number of sites available for selection into a SAT or precipitation proxy network for each region and calibration window size. The range is calculated across the 10 different calibration windows used when creating a network, as discussed in Section 2.3.

3 Results

The importance of a long calibration window is illustrated in Figure 2. For example, a true correlation of -0.3 between precipitation and the SAM may become anything ranging from -0.65 to 0.1 when evaluated over a shorter 31 year window (Figure 2d). However, as the window size increases, it is increasingly likely that the calculated correlation is representative of the true correlation. For example, calibration windows of 61 and 91 years ensure that our proxy's correlation with SAM is always the same sign as over the 500 year period (Figure 2e and f). Also noteworthy is the considerable decrease in the maximum available number of proxies eligible for inclusion in reconstructions when calibrating with a 61 year window, rather than a 31 year window (Table 1). A smaller decrease in the proxy pool is seen when lengthening the window from 61 to 91 years.

3.1 Reconstruction skill

Reconstructions of the SAM often rely heavily on proxies from a very limited geographic location. What follows are several reconstructions, each one utilising proxies from one or more regions which are commonly used to reconstruct the SAM. In the first scenario, pseudoproxies are sourced from the entire Southern Hemisphere (SH), including Antarctica, shown in Figure 3a-f. The reconstruction skill is displayed as a correlation (y-axis) between the pseudoproxy generated SAM index and the

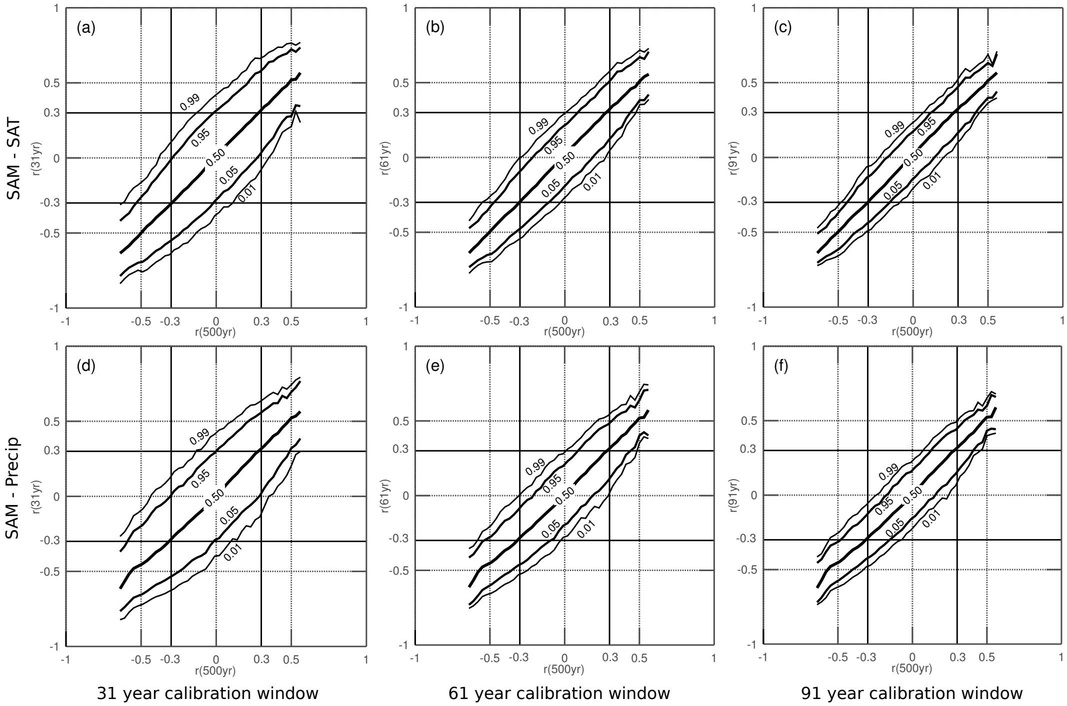

Figure 2. Correlation coefficients between the model SAM index and both SAT (a, b and c) and precipitation fields (d, e and f). Panels show the probability distribution of a gridpoint having a certain probability in a 31 (a and d), 61 (b and e) and 91 (c and f) year calibration window, given that same point's correlation over the full 500 years. This illustrates how a longer calibration window will ensure that the correlation of a point to the SAM within that window will be closer to the 'true' correlation, calculated over the full 500 years.

'real' SAM index calculated from sea level pressure (SLP) fields in the model. This is plotted against the number of proxies used to generate the reconstruction (x-axis) and the range in the 5^{th}, 50^{th} and 95^{th} percentiles represents the use of 10 different calibration windows and the effect this has on the reconstruction.

Results suggest that small proxy networks (2-10 proxies) rarely provide skilful reconstructions of the SAM, even when calibration is a relatively large 91 years (Figures 4 and 5, panels a, e, i; red line), though a greater proportion of precipitation-derived reconstructions are considered skilful across all window sizes. The range in reconstruction skill is smaller for precipitation than for SAT, particularly when longer calibration windows are used, suggesting larger multi-decadal variability in the SAM-SAT teleconnection over time. Maximising the number of records in the proxy network leads to a larger proportion of skilful reconstructions, although for the shortest window of 31 years, the reconstruction skill in the 95^{th} percentile is never greater than $r = 0.76$ for precipitation and $r = 0.77$ for SAT (r^2 = 0.58 and 0.59 respectively), suggesting that at most around 60% of the model SAM variability can be reproduced. Minimum values (lowest r value in the 5^{th} percentile for 70 proxies and 31 year window) are $r = 0.62$ for SAT and $r = 0.59$ precipitation, reconstructing 38% and 35% of SAM variability, respectively (Figure 3a and d).

The range in reconstruction skill presented in Figure 3 indicates that even when the network size is maximised and a long window is selected, simply calibrating during a different period can change the skill of the resulting reconstruction. It is noteworthy that increasing the calibration window length does not necessarily increase the maximum possible skill of the resulting reconstruction, but rather leads to a reconstruction converging towards the skill of a so-called 'true' reconstruction. This 'true' reconstruction utilises the entire time-span of our data for calibration, which is 500 years here, and shows the actual ability of these proxies to reconstruct the SAM. This convergence is visible for the SH SAT reconstructions (Figure 4a, e and i), where a longer calibration window does not increase the 95^{th} percentile of reconstruction skill, nor necessarily increase proportion of skilful reconstructions (Figure 4e and i; red lines). In other words, a longer calibration window will more realistically represent a proxy's relationship with the SAM but as a result, may decrease the 'skill' of the reconstructed SAM.

As Antarctica represents a large percentage of the available proxies (Table 1), reconstructions are included for proxies sourced from the entire Southern Hemisphere other than Antarctica to ensure they are not disproportionately impacting the skill of our reconstructions. The Antarctic-free SAT reconstructions are less skilful for the 31 and 61 year windows with a larger range in r. Of note is that most of the 95^{th} percentile (Figure 3g, h, i - red shading) is below the $r^2 > 0.5$ skilful threshold, as opposed to reconstructions with Antarctic sites. Antarctic-free precipitation reconstructions typically see an increase in maximum skill but a similar increase in the range (Figure 3i-l). The contrasting effects of Antarctica could be due to Antarctic precipitation having a generally weak correlation with the model SAM, while SAT shows strong negative correlation with the SAM continent-wide (Figure 1) and by removing these points, we lose skill in the SAT-derived reconstructions and increase skill in the precipitation-derived reconstructions.

Data from different regions may also act to increase or decrease the skill of reconstructions. Figures 4 and 5 illustrate the skill of each regional reconstruction in comparison to the SH one. In addition, comparisons are made to a reconstruction with a 'true' calibration window of 500 years, showing the actual range in skill that the pseudoproxies can produce. Southern Africa was excluded from this analysis as too few grid cells met our criteria for reconstruction. When we utilise records from individual regions the reconstructive skill of the proxy network is significantly reduced. Reconstructions for the Australia-New Zealand region (Figure 4 and 5c, g, k), South America (Figure 4 and 5b, f, j) and Antarctica (Figure 4 and 5d, h, l) all show reduced reconstructive skill when compared with the entire SH network (Figures 4 and 5a, e, i), with Antarctica being the only individual region capable of generating any skilful reconstructions. In general then, reconstructing the SAM using pseudo-proxies in CM2.1 is most successful when we maximise network size and source sites from as many geographical regions as possible, particularly at longer calibration windows, where the proxy pool becomes too small for a full network in many regions. The exceptions here are precipitation-based reconstructions, where leaving out Antarctica improves reconstruction skill.

When comparing proxy types, there are significantly more skilful precipitation-derived reconstructions than for SAT, and this is true across all window sizes for the Southern Hemisphere-wide reconstructions (Figures 4 and 5a, e, i). In particular, when using a 61 or 91 year window, 89% and 91% of SH precipitation reconstructions are considered skilful, respectively (with a network size of 70)(Figure 5e and i) and an increase in window size increases the proportion of skilful reconstructions (Figure 5a, e, i; red line). In contrast, SAT reconstructions calibrated with 61 and 91 year windows only produce skilful

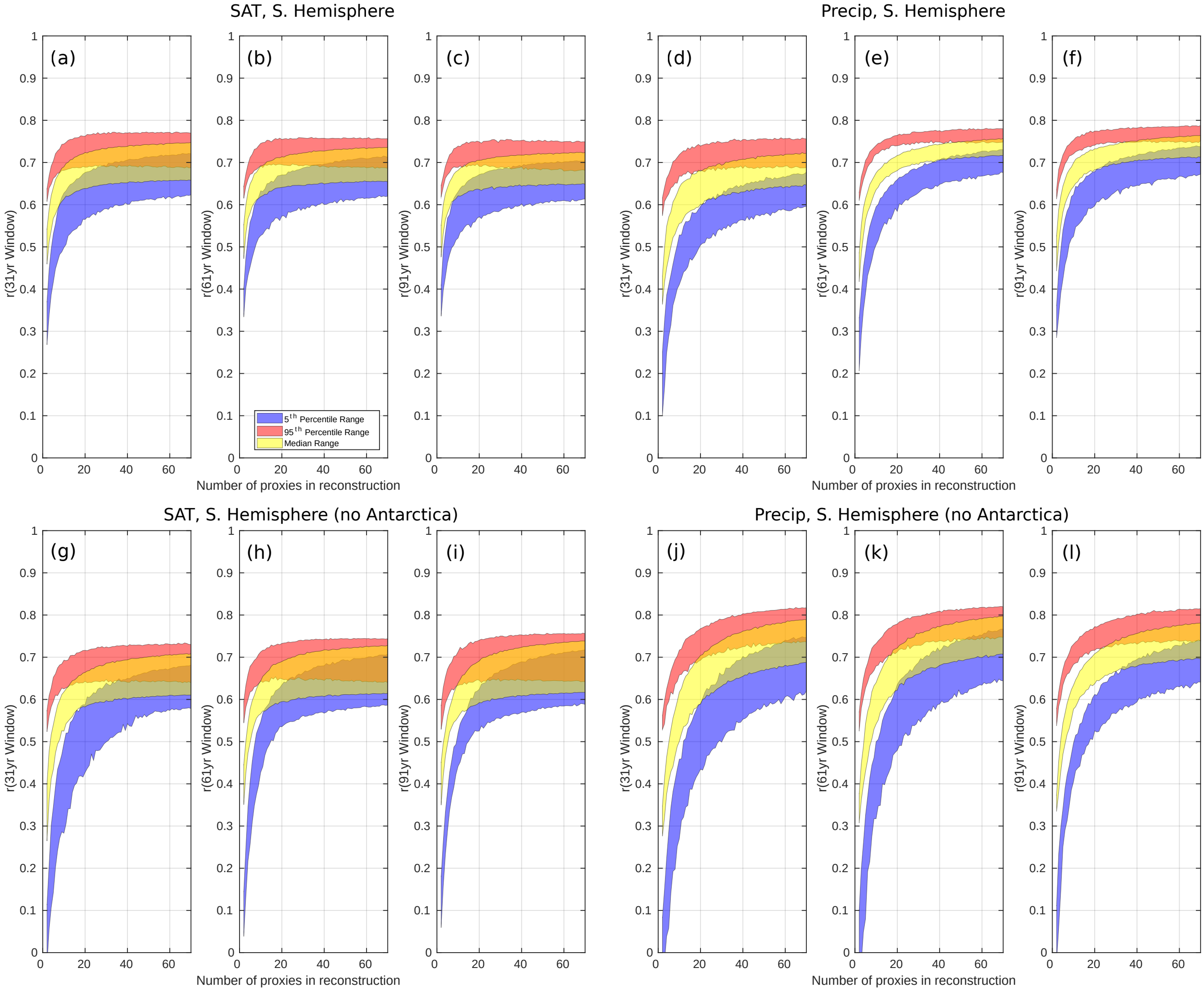

Figure 3. Correlation of the model SAM index (y-axis) to the pseudo-proxy reconstructions described in Section 2.3, plotted here by network size (x-axis). Each panel shows reconstruction skill for the 31, 61 or 91 year calibration windows. The three shaded areas show the 5^{th}, 50^{th} and 95^{th} percentile. Their range represents the range of these percentiles across the 10 different calibration windows for each window/ network size. Panels a-c and d-f show reconstructions for SAT and precipitation respectively, where proxies are sourced from the entire Southern Hemisphere. Panels g-i and j-l show reconstructions generated by proxies sourced from everywhere but Antarctica, but are otherwise equivalent to a-f.

reconstructions 25% and 20% of the time, respectively (Figure 4e and i). Most striking here is that a longer calibration window both decreases the 95^{th} percentile skill and the proportion of SAT-derived reconstructions that can be considered skilful. But this is reasonable when we see that, at best, 11% of 'true' SAT reconstructions are skilful and have a lower maximum skill for the 95^{th} percentile (Figure 4e and i). This result indicates that shorter calibration windows are sufficiently susceptible to

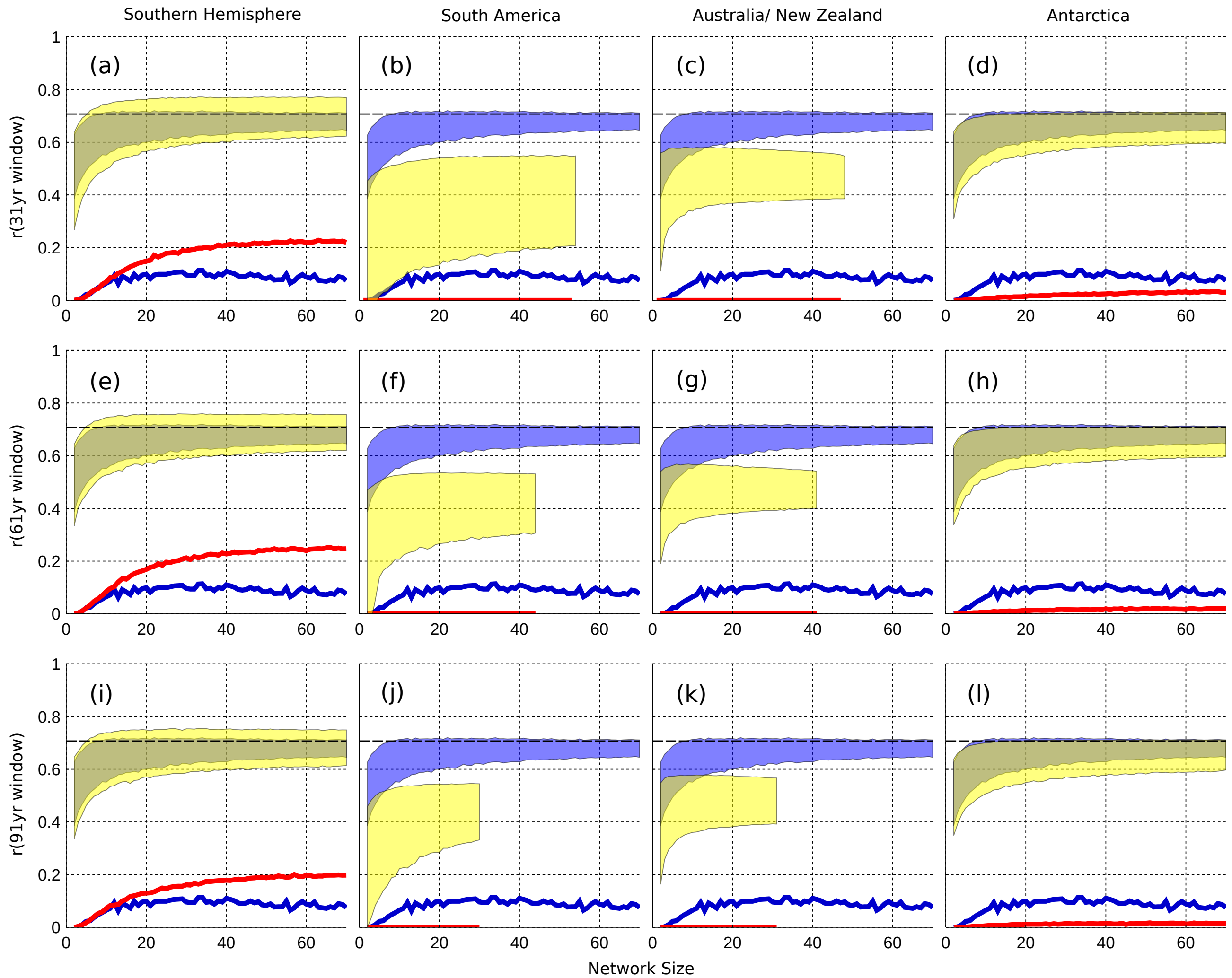

Figure 4. Differing reconstruction skill achieved when using SAT-derived proxies sourced from the entire Southern Hemisphere (a, e, i), South America (b, f, j), Australia and New Zealand (c, g, k) and Antarctica (d, h, l) only. The correlation between a SAT-derived reconstruction and the SAM is on the y-axis, while the number of sites used (n = 2:70) in a reconstruction is on the x-axis. Shaded regions represent the range between the minimum of the 5^{th} percentile and the maximum of the 95^{th} percentile for each network size, across 10,000 reconstructions (described in Section 2.3). Each set of regional reconstructions is shaded in yellow and the end of this yellow region indicates the number of samples available when it is below 70. Each panel also includes the range in skill for reconstructions with sites sourced from the entire Southern Hemisphere and calibrated with an 'true' 500 year window (blue shading). The blue line indicates the percentage of 'true' SH reconstructions that meet or exceed our skill threshold of being able to explain 50% or more of the variability in the SAM. The red line indicates the same thing, but for each regional reconstruction. The dashed black line indicates the r value required to meet our skill threshold.

climatic noise or modulation that they are producing reconstructions with spuriously larger reconstruction skill. It is also worth noting that the reduction of reconstruction skill range visible for the 61 and 91 year windows relative to the 31 year window

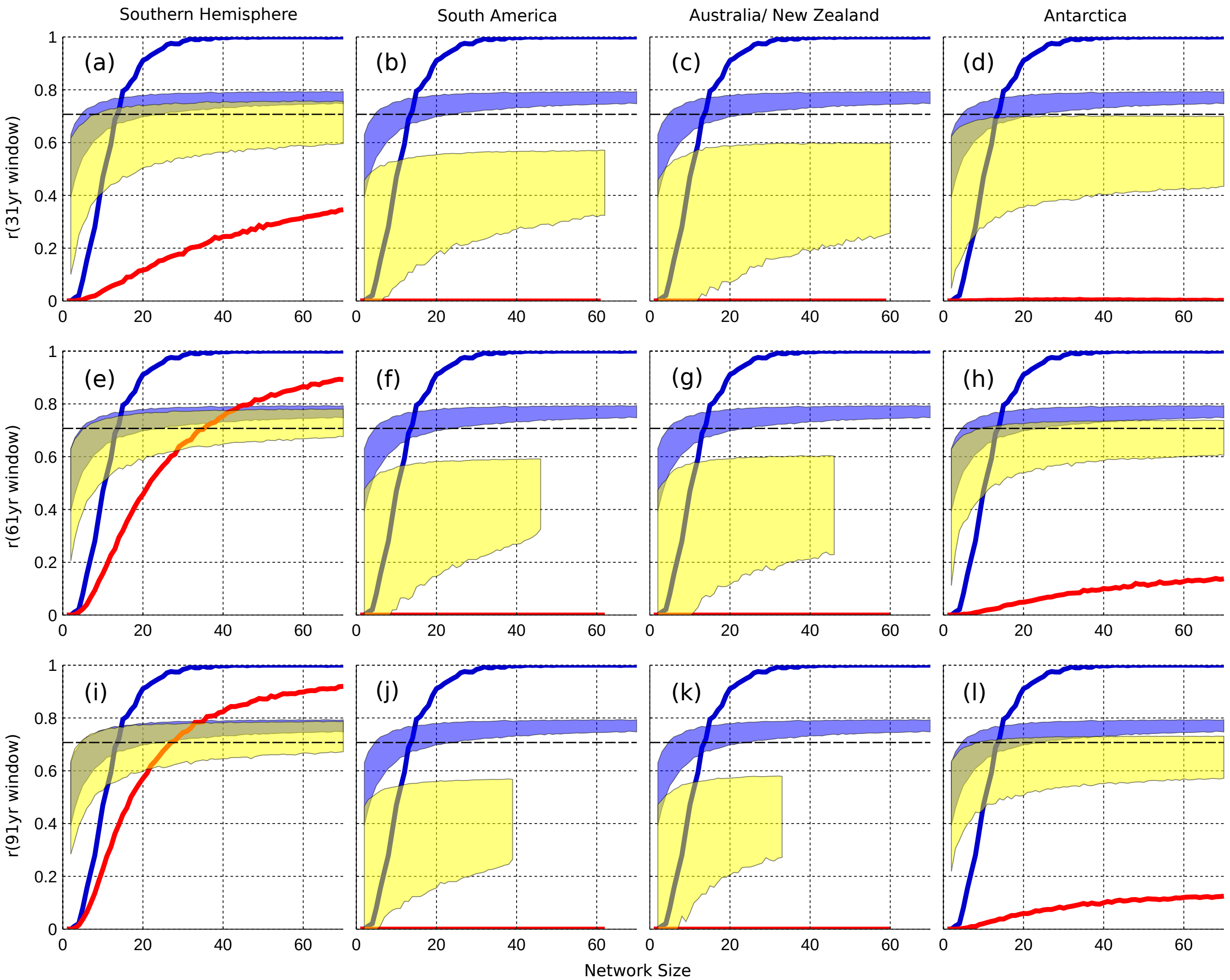

Figure 5. Differing reconstruction skill achieved when using precipitation-derived proxies sourced from the entire Southern Hemisphere (a, e, i), South America (b, f, j), Australia and New Zealand (c, g, k) and Antarctica (d, h, l)) only. The correlation between a precipitation-derived reconstruction and the SAM is on the y-axis, while the number of sites used (n = 2:70) in a reconstruction is on the x-axis. Shaded regions represent the range between the minimum of the 5^{th} percentile and the maximum of the 95^{th} percentile for each network size, across 10,000 reconstructions (described in Section 2.3). Each set of regional reconstructions is shaded in yellow and the end of this yellow region indicates the number of samples available when it is below 70. Each panel also includes the range in skill for reconstructions with sites sourced from the entire Southern Hemisphere and calibrated with an 'true' 500 year window (blue shading). The blue line indicates the percentage of 'true' SH reconstructions that meet or exceed our skill threshold of being able to explain 50% or more of the variability in the SAM. The red line indicates the same thing, but for each regional reconstruction. The dashed black line indicates the r value required to meet our skill threshold.

will necessarily be in part due to the overlapping of the 10 calibration windows over the 500 years of model data. With a longer data set, the lack of such an overlap would almost certainly result in this spread being larger.

While increasing the number of sites used in each reconstruction does not necessarily improve the maximum 95^{th} percentile skill after approximately n = 10, it does narrow the range of possible reconstruction skill (Figures 4 and 5a, e, i; note the yellow envelope converging on the blue with increasing window size). While Figures 4(a,e,i) give the impression of our reconstructions outperforming the 'true' reconstructions, they have virtually the same maximum skill at each network size. This apparent incongruence occurs due to the probability distribution of reconstruction skill for our 'true' proxies being far narrower than for the reconstructions with varying window length, resulting in the 95^{th} percentile having a generally lower value for each network size.

3.2 Mapping non-stationarity

In this section we examine whether certain regions are more or less non-stationary to the SAM, which would contribute to these regions being better or worse than others at reconstructing the SAM. Figure 6 shows the number of non-stationary years at each grid point for SAT (a) and precipitation (b) as defined in Section 2.2. Grid points with running correlations that fall outside the 95% confidence interval of stochastic variability more than 10% of the time are highlighted with solid contours; we define these regions as non-stationary. For SAT, 4% (31 year window) and 8% (61 and 91 year window) of land cells are non-stationary, while for precipitation 6% (31 year window, 10% (61 year window), and 9% are (91 year window).

Depending on the length of the calibration window, different patterns of non-stationarity appear, particularly for SAT. Aside from three small regions in south-east Australia, central South America and the Queen Elizabeth Range in Antarctica, there are almost no land sites that can be considered non-stationary when using a 31 year running correlation for SAT. It is also noteworthy that these non-stationary regions (as calculated using the 31 year running correlation) appear to fall, on average, in regions where correlations are weaker (though still significant at $p < 0.1$ when $r > 0.08$) over the full 500 years (Figure 1a). The same is also broadly true for precipitation, where large regions of non-stationary points do occur but fall in regions of weaker or 0 correlation with the SAM, particularly in East Antarctica (Figure 1b). It is worth noting, however, that despite not meeting the requirement of being classified as non-stationary, large regions of the Southern Hemisphere land surface show modulation of the SAM-proxy teleconnection (Figure 6, yellow regions).

To better illustrate the impact of non-stationary proxies on reconstructions, Figure 7(a) compares the skill of our SH reconstructions with the percentage of non-stationary proxies in each. The effect of non-stationary sites is negative in all but one instance. Correlations are typically stable with network size and are relatively weak, with mean r^2 values of 0.03. Reconstructions calibrated with a 31 year window are outliers, both of which see a slight increase in skill with larger network sizes. In particular, the positive relationship observed for the precipitation reconstructions (Figure 7a and b, purple line) suggests that these proxies provide a net benefit to the reconstructions they are part of, despite their non-stationary nature. SAT reconstructions calibrated over 61 and 91 years are noteworthy as the impact of non-stationary sites is larger (r^2 = 0.19 for 70 proxies calibrated over 91 years) and increases with network size, when compared to other scenarios (Figure 7a and b, yellow line).

The negative relationship between reconstruction skill and non-stationarity belies the chances of producing a reconstruction with a large proportion of non-stationary proxies. Figure A1(c and d) demonstrates that, for a network size of 70, the most likely proportion of non-stationary proxies in a reconstruction is ∼5%, and even this only constitutes 10-15% of reconstructions.

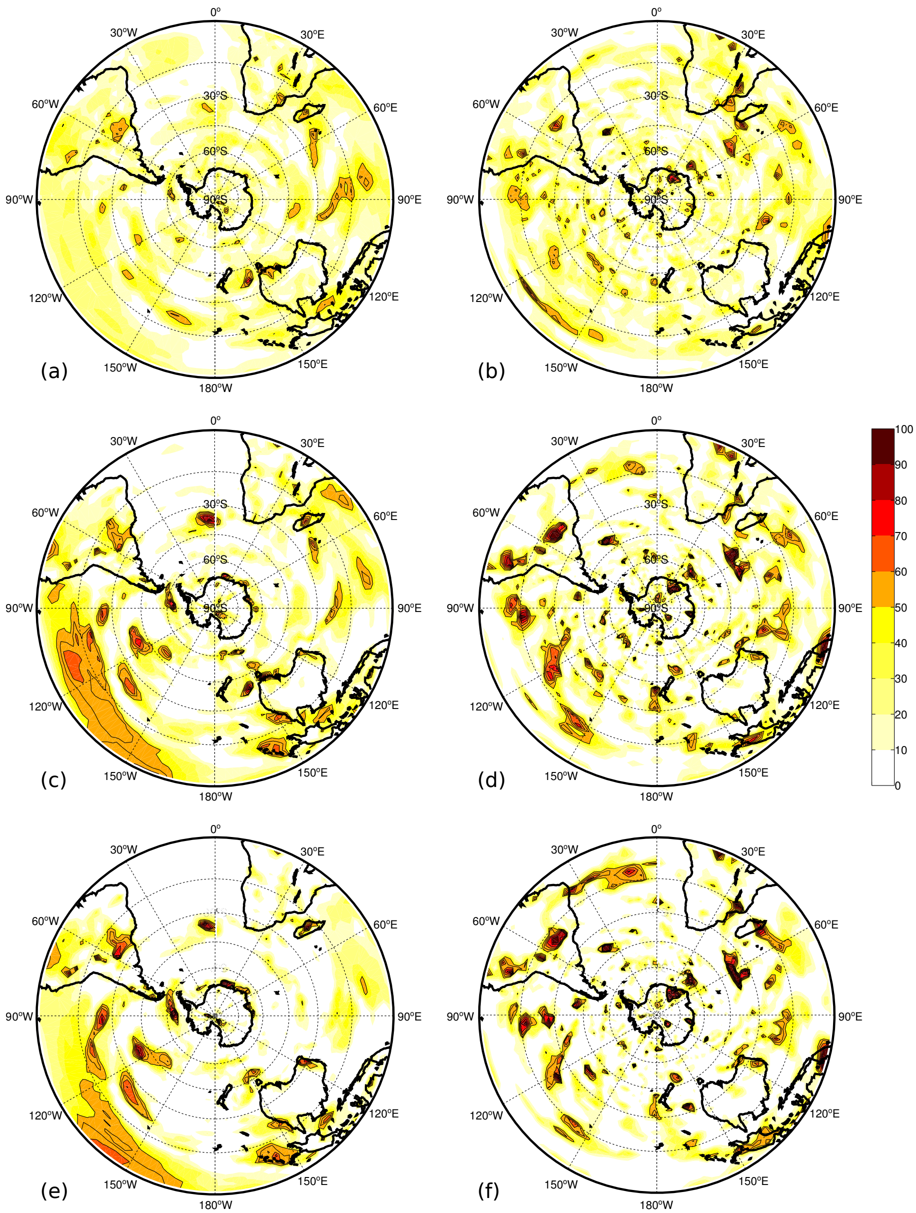

Figure 6. Number of years at each gridpoint where the 31 year (a,b), 61 year (c,d) and 91 year (e,f) running correlation between SAT (a,c,e) or precipitation (b,d,f) and the model SAM falls outside the 95% 'stationarity' confidence interval (Section 2.2). As per our definition of non-stationarity, regions which fall outside this interval 10% of the time or more (≥ 50 years) are highlighted with solid black contours and are considered to be non-stationary.

While increasing the calibration window results in a larger number of non-stationary sites in a reconstruction, a sufficiently large proxy network minimises the probability that these non-stationary sites will represent a significant proportion of the network. In summary, there is a weak negative relationship between the proportion of non-stationary proxies in a reconstruction and its skill, but this impact is not felt by the majority of our reconstructions.

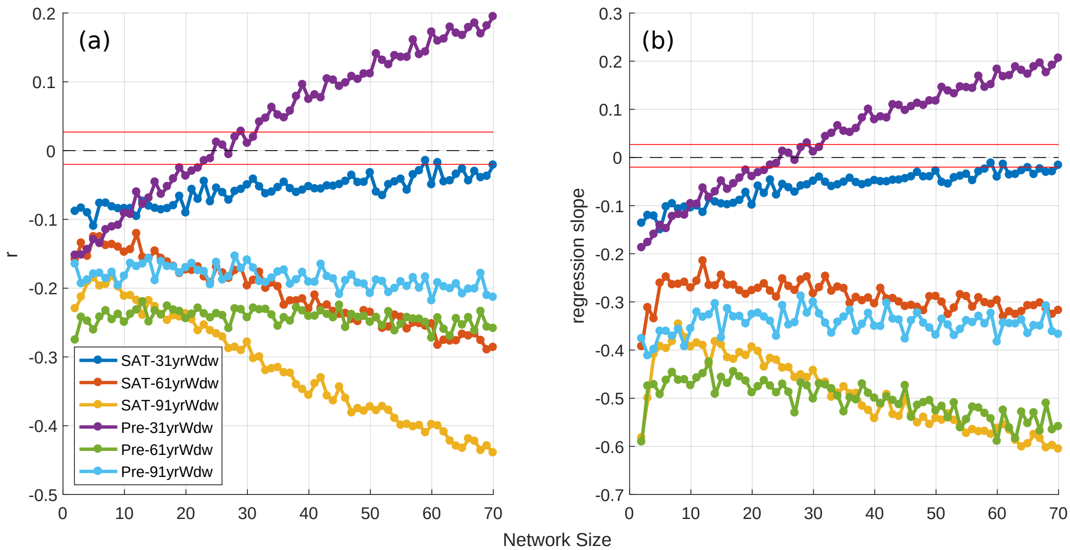

Figure 7. Correlation (y-axis) between the skill of a given reconstruction and the percentage of non-stationary proxies it contains (a), plotted as a function of network size. (b) is the same as panel (a), but y-axis shows regression slope. Calculations are over 10,000 reconstructions for each network size. $r = 0$ is plotted as a black dashed line. All correlations are significant to at least $p < 0.05$ other than in the region bounded by the two red lines about $r = 0$.

3.3 Modulation of the SAM-proxy teleconnection

While few terrestrial cells qualify as non-stationary based on the definition in Section 2.2, there is still considerable variance in the teleconnection strength between SAM and SAT/precipitation over the 500 years of the simulation (Figure 8). While this could be due to climatic noise, it is not unreasonable that other modes of climatic variability - in particular, ENSO - may be modulating this teleconnection (Silvestri and Vera, 2009; Fogt et al., 2011; Dätwyler et al., 2020). The regions from which we source our proxies, such as Australia/ New Zealand and South America are strongly impacted by ENSO, with its teleconnections visible in both temperature and precipitation fields (Davey et al., 2014). The following section will examine which regions show the most variance in proxy-SAM teleconnection and whether or not these regions appear to be influenced by the model ENSO.

Variations in SAM teleconnection strength for SAT proxies shows considerable variance in Antarctica, northern and southern South America and for the 31 and 61 year windows, parts of Australia and New Zealand (Figure 8a-c). Distinct regions of higher teleconnection variance in Antarctica are typically in regions of high SAM-SAT correlation over the full 500 years (Figure 8a-c; dashed contours), with this variance decreasing as we move to a 91 year calibration window. Further to this, variance is typically low for regions with a small 500 year r value.

Significant correlations between the running correlations of SAM-SAT and the filtered n3.4 index can be seen over much of Western Antarctica, south-eastern Australia and parts of South America (all windows, Figure 9,a-c), while significant correlations are also seen in East Antarctica in the 31 year window. ENSO's modulating influence can be seen to vary, depending on both the strength of the underlying SAM-proxy correlation as well as the calibration window length (Figure 11). For the 31 year window, the regression coefficient between ENSO and the SAM-SAT running correlation is relatively low, clustering predominately between 0.2 and 0.4 (Figure 11a), and generally decreases if a site has a stronger correlation with SAM over the full 500 year period. A similar relationship is visible for the 61 year window, however the regression coefficients are slightly larger ($\sim 0.25 - 0.5$) and the decrease in ENSO regression coefficient with increase in SAM-SAT r is less apparent. The 91 year window sees this relationship disappear altogether, with relatively large ($\sim 0.6 - 0.8$) regression coefficients independent of SAM-proxy correlation (Figure 11c and f). This is of lesser consequence, however, as most sites show little variance in SAM-SAT teleconnection at this longer window length (Figure 8c; most regions have an $r_{std} < 0.1$) despite ENSO potentially being responsible for 50% of this variance. Furthermore, any impact of ENSO on the SAM-SAT telconnection can be reduced with a longer calibration window as the number of land points (SAM-SAT running correlation) significantly correlated with the filtered n3.4 index decreases as the window length increases (i.e., this is respectively 30%, 24% and 15% for the 31, 61 and 91 year windows).

For precipitation, teleconnection strength is less variable and only parts of Australia, Indonesia and the Ross Ice Shelf/ Marie Byrd Land in Antarctica show large changes (Figure 8d-f). Correlation of this precipitation teleconnection variance with the model n3.4 index reveals few regions of significant ENSO influence (Figure 10) and little coherent spatial structure for this correlation. The magnitude of the impact of ENSO on the SAM-precip teleconnection is similar to that of SAT proxies (Figure

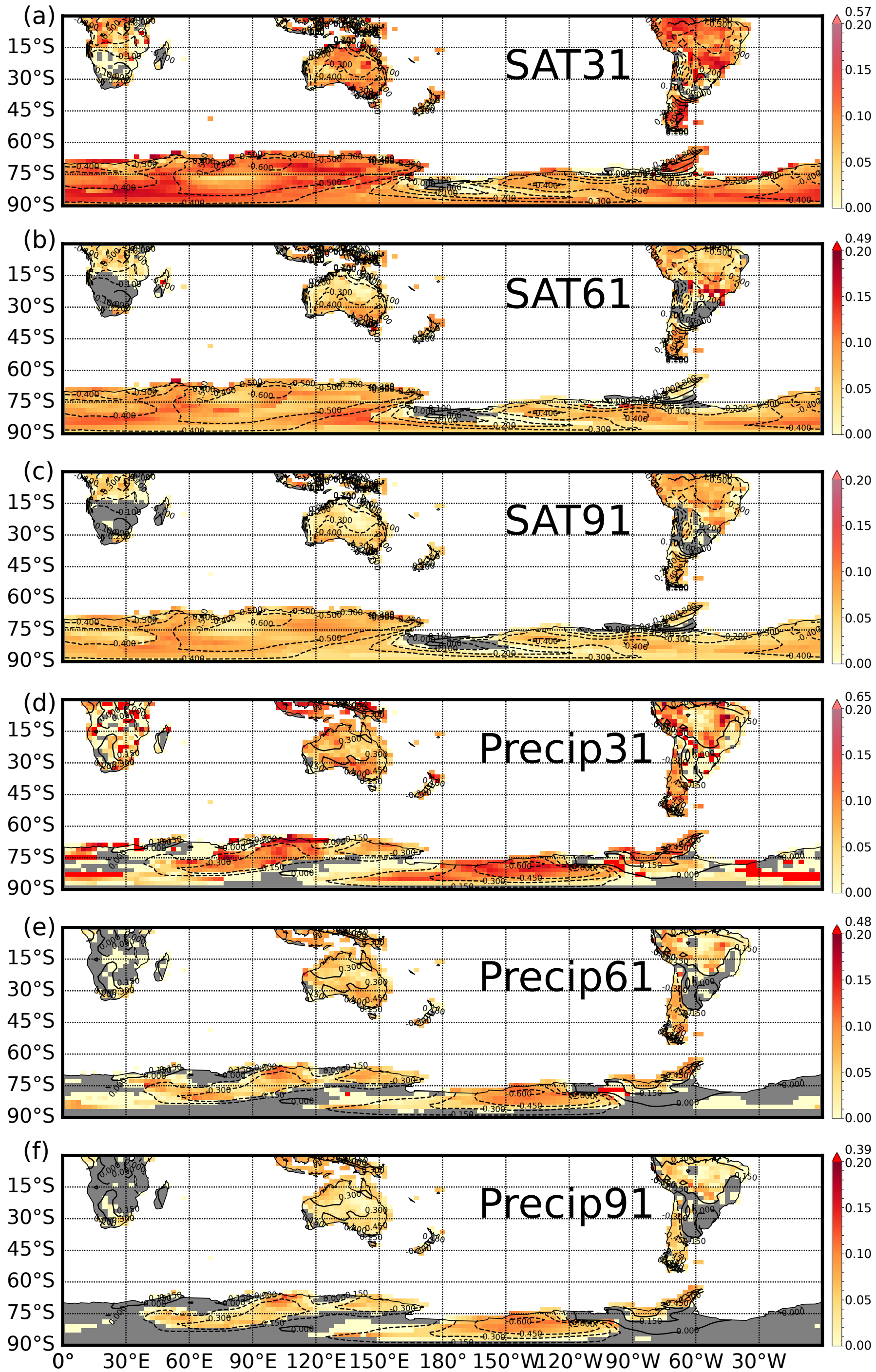

Figure 8. One standard deviation of correlations between SAT (panels a-c) and precipitation (panels d-f) with the model SAM index, over the 10 calibration windows. Panel a) shows values for the 10, 31 year calibration windows for SAT. Panels (b) and (c) are for the 61 and 91 year windows respectively. Panels (d), (e) and f) are the same, but for precipitation. Maximum values on colour bars for panels (a),(b),(d),(e) and (f) indicate the colour of several outliers in the data. Grey regions indicate cells that did not meet the minimum correlation criteria (r ≥ |0.3|) over 8 or more windows, meaning no standard deviation could be calculated. Dashed contours show the model SAM-SAT (a-c) and model SAM-precipitation (d-f) 500 year correlation fields from Figure 1.

11d, e, f). The number of grid cells impacted by ENSO is fewer than that for SAT, with the running correlation of SAM-precip being significantly correlated with n3.4 in 21%, 16% and 8% of land cells for the 31, 61 and 91 year windows.

Removing these ENSO sensitive proxies from our SH-wide reconstructions has a small, but negative impact on the proportion of skilful reconstructions we are able to produce for both SAT and precipitation (Figure 12). Their absence also reduces the minimum skill values for the 5^{th} percentile for all precipitation-derived reconstructions across all network sizes (Figure 12d, e, f). A smaller effect is visible for SAT-derived reconstructions calibrated with a 31 year window, but only for smaller network sizes. Given the minimal extent to which ENSO appears to modulate the proxy-SAM relationship, removing these proxies, which may otherwise enhance the regional diversity of a network, results in a net degradation of the signal to noise ratio in our reconstructions. On the other hand, reconstructions using only ENSO-sensitive proxies (not shown) also results in lower skill, although it is unclear what role ENSO plays due to the vastly reduced pool of proxies we can sample from in this scenario.

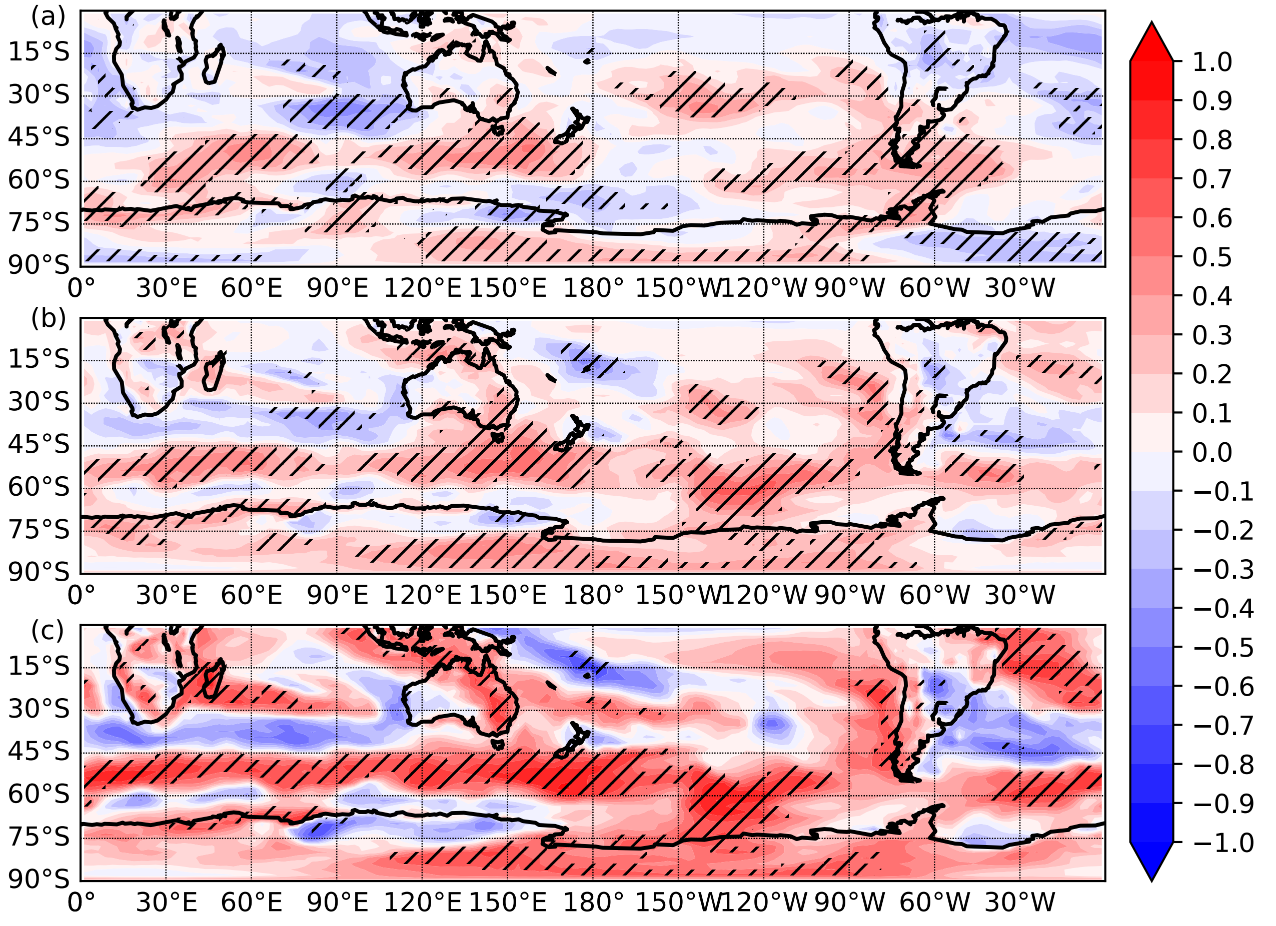

Figure 9. Correlation between the (a) 31, (b) 61 and (c) 91 year SAM-SAT running correlation at each grid cell and the model-derived n3.4 index. The n3.4 index is filtered with a corresponding 30, 60 and 90 year filter. Hatched regions indicate p < 0.05.

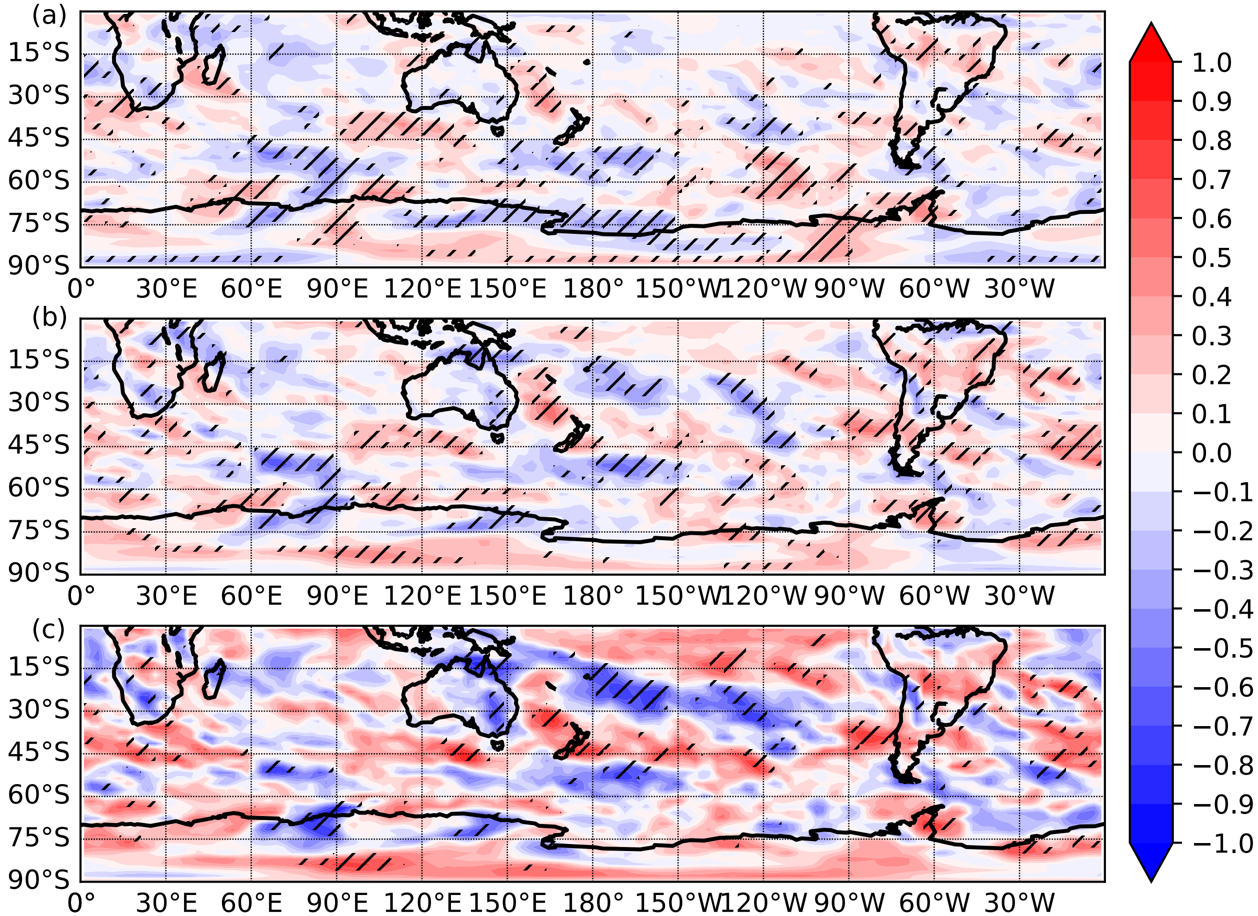

Figure 10. Correlation between the (a) 31, (b) 61 and (c) 91 year SAM-precipitation running correlation at each grid cell and the model-derived n3.4 index. The n3.4 index is filtered with a corresponding 30, 60 and 90 year filter. Hatched regions indicate $p < 0.05$.

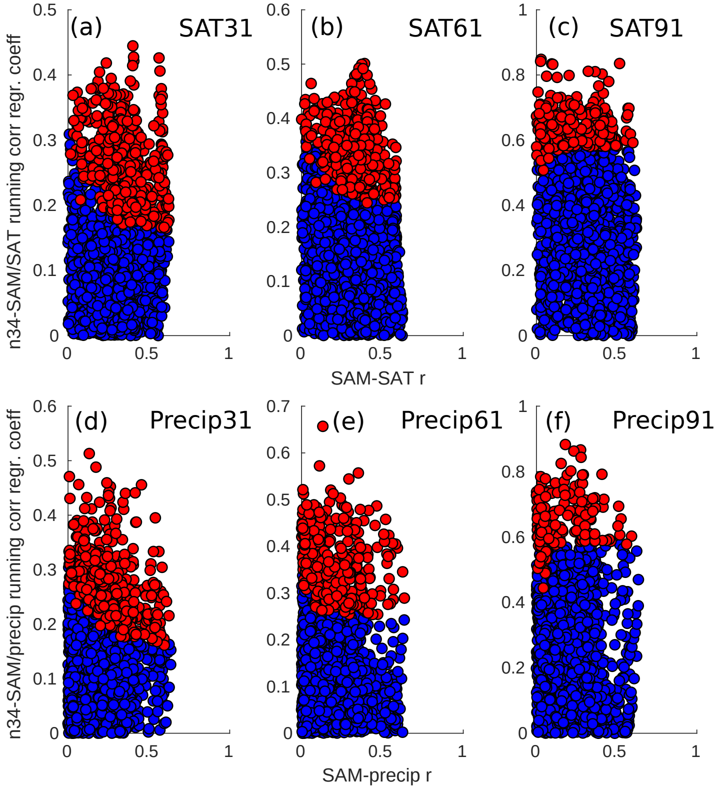

Figure 11. Scatter plot of the regression slopes from the significant (red) and non-significant (blue) land points shown in Figures 9 and 10 plotted against the 500 year correlation coefficient between the SAM and SAT (a, b, c) and precipitation (d, e, f). Both SAM-proxy running correlations for each cell and the n3.4 index are standardised prior to regression calculation. Note the changing vertical scales between panels.

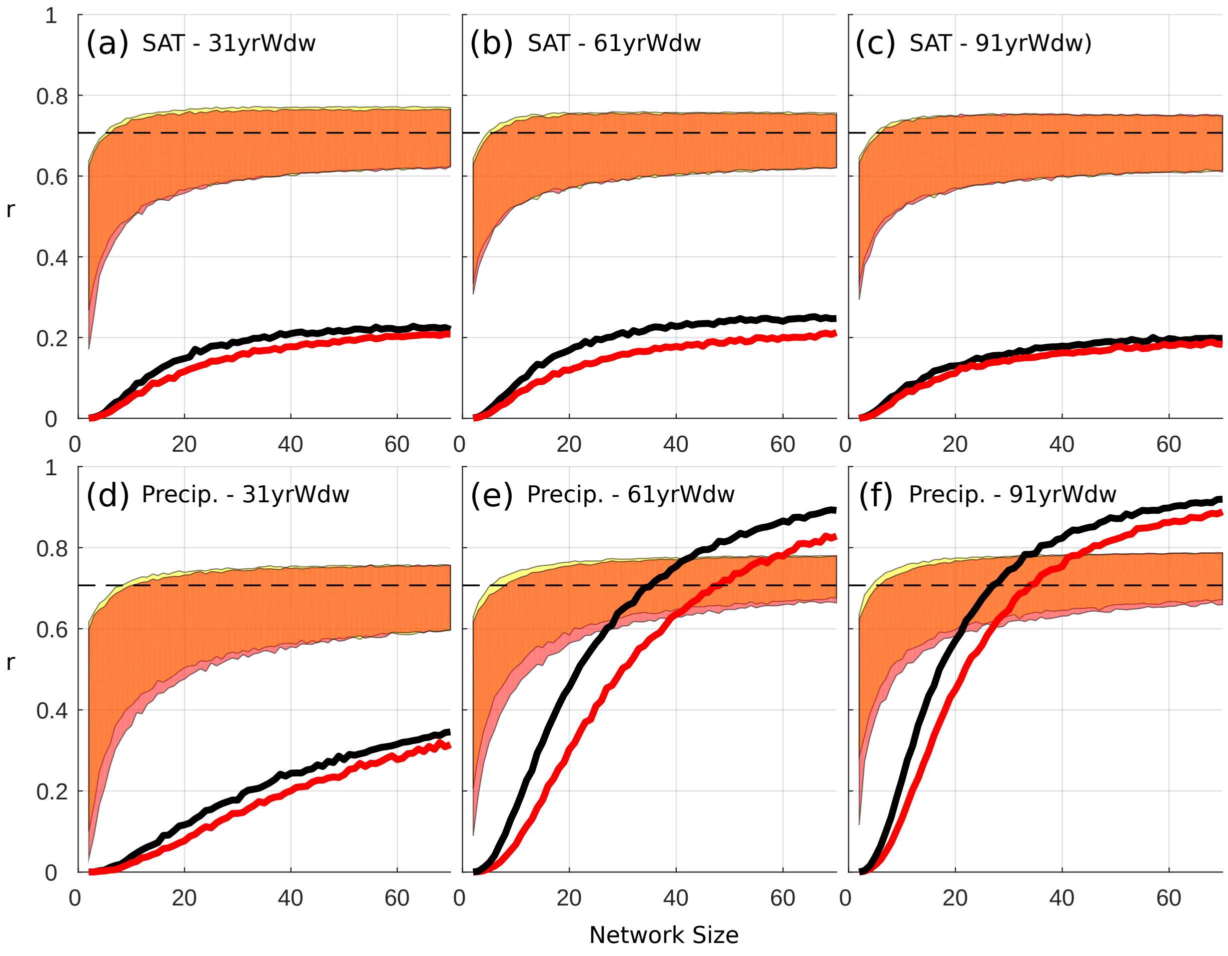

Figure 12. Differing reconstruction skill achieved when sourcing proxies from the entire Southern Hemisphere (yellow envelopes) and the entire Southern Hemisphere, excluding proxies whose teleconnection with SAM have a significant ($p < 0.05$) correlation with ENSO (hatched regions in Figures 9 and 10; red envelope). The correlation between a SAT or precipitation-derived reconstruction and the SAM is on the y-axis, while the number of sites used (n = 2:70) in a reconstruction is on the x-axis. Shaded regions represent the range between the minimum of the 5^{th} percentile and the maximum of the 95^{th} percentile for each network size, across 10,000 reconstructions (described in Section 2.3). The black lines indicate the percentage of SH reconstructions (yellow envelope) that meet or exceed our skill threshold of being able to explain 50% or more of the variability in the SAM. The red line indicates the same thing, but for those reconstructions that exclude ENSO-sensitive proxies (red envelope). The dashed black line indicates the r value required to meet our skill threshold.

4 Discussion and Conclusions

In this study we use the CM2.1 coupled climate model to examine limits to SAM reconstruction skill, including the impact of regional biases in sourcing of proxy records, as well as the impact of non-stationary proxy teleconnections. Reconstructions derived from model SAT and precipitation fields and calibrated over a 31 year window are able to - at best - replicate 56% to 58% of SAM variance respectively, comparing favourably to a 'true' reconstruction whose proxies are calibrated over the full 500 year interval (Figures 4 and 5). This suggests a possible upper limit to the variance an annual-mean SAM reconstruction can reproduce of $\sim$60%.

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

Both SAT and precipitation-derived reconstructions were most skilful when proxies were selected from a geographically broad region, while regional reconstructions - with the exception of Antarctica - fail to produce any skilful reconstructions. This is likely due to each region being affected by localised climatic noise which becomes a systematic source of error in the reconstruction. With larger datasets from different regions, this noise cancels out and the signal we seek to reconstruct is more clearly visible.

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

Low frequency variability exists in the teleconnections between our pseudoproxies and the SAM that cannot be explained by climatic noise (stochastic variability). CM2.1 simulates, at maximum, 9% of land points as being non-stationary as defined by Gallant et al. (2013) (using precipitation as a proxy and a 91 year running window), although the odds of creating a proxy network with a high proportion of non-stationary sites remains relatively low (Figure A1). Non-stationary proxies, as defined here, do not seem to modulate SAM-proxy teleconnection strengths or impact on reconstructions greatly, as emphasised by the weak relationship between reconstruction skill and the number of non-stationary proxies in a reconstruction (Figure 7a).

The exceptions are SAT-derived reconstructions with a longer calibration window (61 or 91 years), suggesting that at larger network sizes, care should be taken to minimise the proportion of non-stationary proxies. While assessing the stationarity of a proxy-SAM correlation is more difficult in the real world, we suggest that multiple methods be employed where possible such as in Dätwyler et al. (2018).

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

As we use only one integration from a single model, it is worth discussing the performance of CM2.1. Its representation of the SAM is good with respect to similar models (Karpechko et al., 2009; Marshall and Bracegirdle, 2015; Bracegirdle et al., 2020), though it does have some biases which may impact the results presented here. For instance, when compared to observations and reanalysis, there is a small equatorward bias in the Southern Hemisphere westerlies in CM2.1, however, the spatial structure and amplitude of SLP anomalies associated with these winds, and therefore the SAM, are well simulated (Delworth et al., 2006). These comparisons are encouraging, particularly considering the multi-decadal changes in the teleconnections that have been observed from in-situ temperature and precipitation measurements (see Silvestri and Vera (2009), Figure 1; Gillet et al. (2006), Figure 1). Furthermore, we may expect that the model SAM's expression in SAT and precipitation fields in our control simulation may not be identical to that found in observations, given the significant positive trend in the SAM over the last 5 decades and potentially accounting for some of the model-data discrepancy.

Assessing these results in individual seasons was beyond the scope of this study, in part due to the CMIP5 generation of models (including CM2.1) being less skilful at representing seasonal variability in the SAM-SAT relationship than over the annual-mean (Marshall and Bracegirdle, 2015), in addition to the hope that reconstructing the SAM on an annual-mean time-scale would smooth out high-frequency noise and enhance the signal to noise ratio of our reconstructions.

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

Most importantly, our results confirm that calibrations of paleo-data to instrumental records over brief time periods can result in misleading teleconnection strengths. The large range in reconstruction skill due to the use of multiple calibration windows suggests real life proxy records may provide a misleading representation of reconstructed SAM variability due to this non-stationary behaviour, particularly when the reconstruction network constitutes fewer than 20 proxies. For a SAT-derived, Southern Hemisphere wide reconstruction, a longer calibration window minimises this uncertainty, but doesn't necessarily result in a more skilful reconstruction. Rather, the reconstructions converge on the 'true' skill that these proxies provide. The use of teleconnection stability screening as utilised in Dätwyler et al. (2018) is an important step in the right direction, and should be utilised alongside correlation 'skill' scores and other validation statistics to assess the reliability of a reconstruction. The lack of long term observational data makes it difficult to circumvent this problem, however climate models which have demonstrated realistic dynamical mechanisms may aid us in calculating the uncertainty of these calibrations in the future.

Code and data availability. TEXT

Model data was downloaded from ftp://nomads.gfdl.noaa.gov/gfdl_cm2_1/CM2.1U_Control-1860_D4/pp/, last accessed 30/08/20. The Marshall SAM index was downloaded from http://www.nerc-bas.ac.uk/public/icd/gjma/newsam.1957.2007.seas.txt, last accessed 9/10/20. ERA-Interim data was downloaded from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim, last accessed 9/10/20. Code for analysis and plotting of figures can be found in the following Github repository: https://github.com/whuiskamp/SAM_pseudoproxy

Analysis and plotting was done using Matlab v2017b, Pyferret v7.4 (PyFerret is a product of NOAA's Pacific Marine Environmental Laboratory. http://ferret.pmel.noaa.gov/Ferret/) and python3.

Appendix A

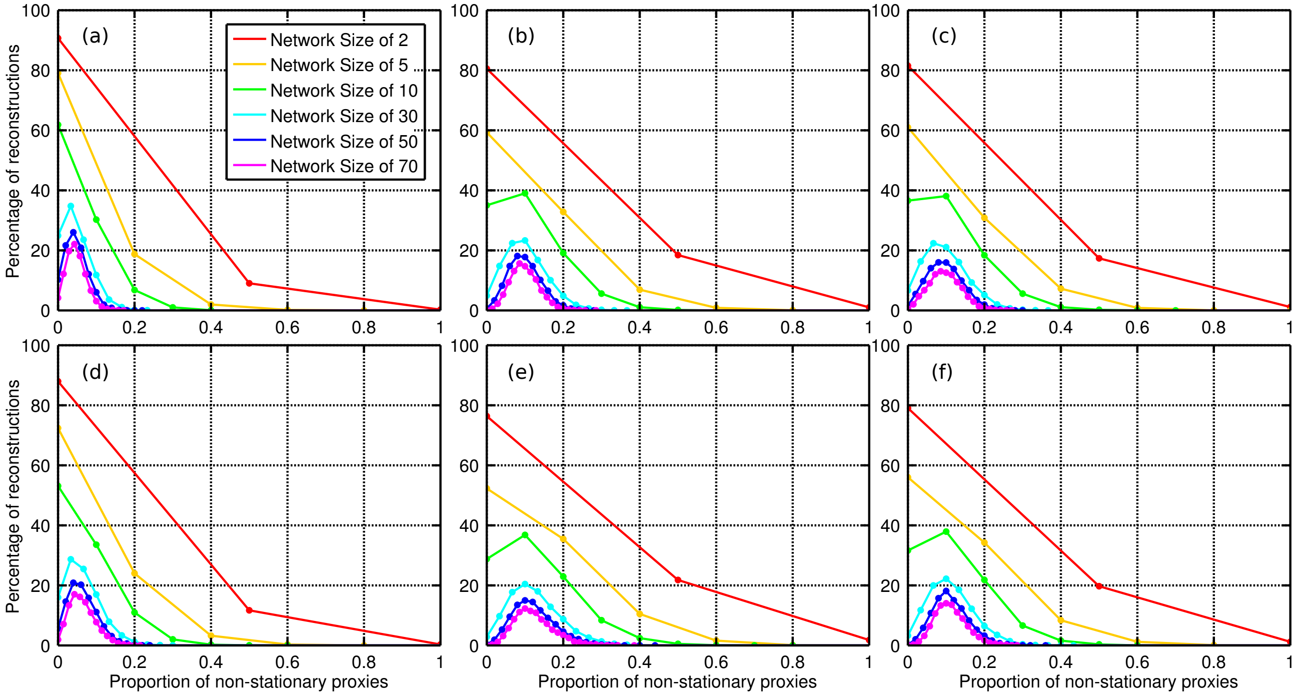

Figure A1. The chance of creating a SAM reconstruction (y-axis) with a certain proportion of non-stationary proxies (x-axis) as calculated from 10,000 reconstructions for each network size. Panels show probabilities for 31 (a and d), 61 (b and e) and 91 (c and f) year calibration windows. Panels a-c and d-f show reconstructions based on SAT and precipitation data respectively.

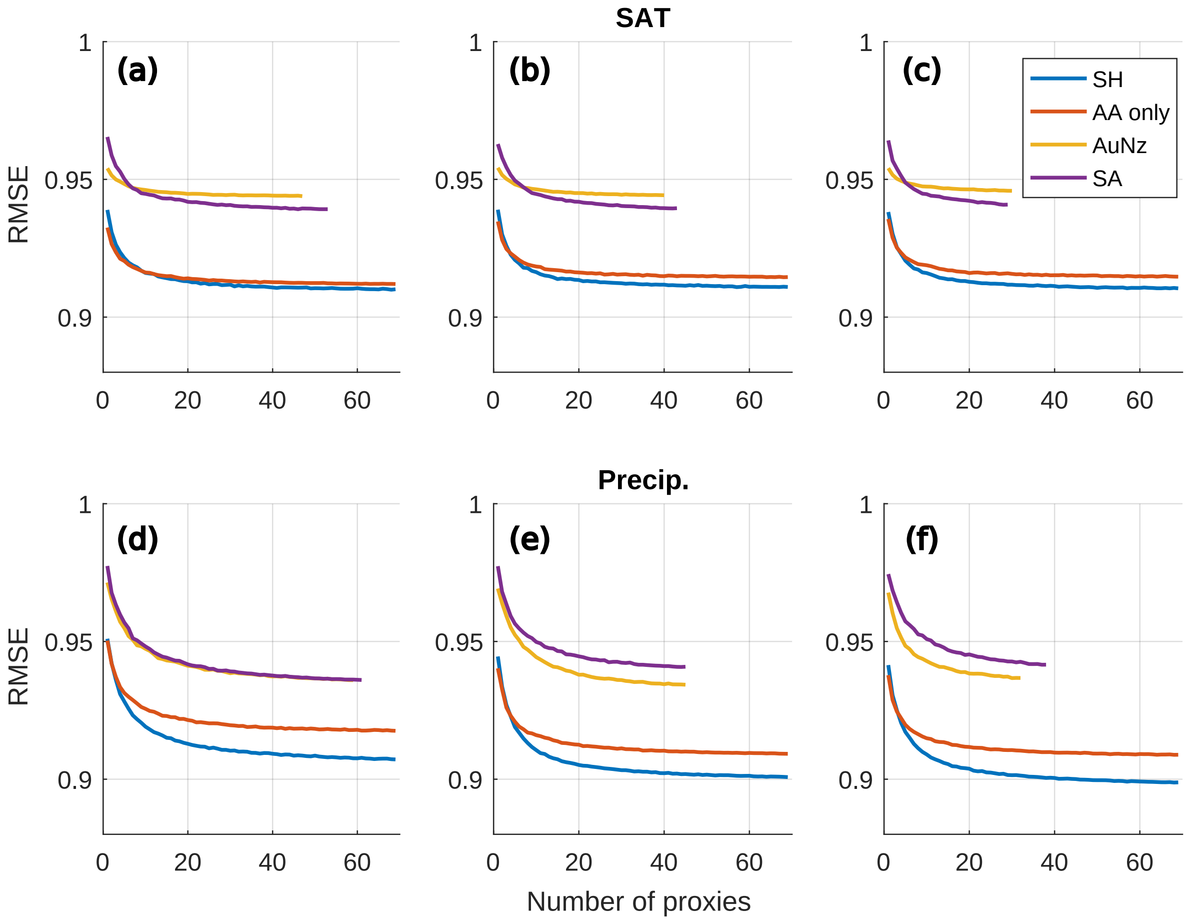

Figure A2. Median root mean square error across the 10,000 reconstructions calculated for each network size for SAT (a, b, c) and precipitation-derived reconstructions (d, e, f). Results are displayed for the 31 (a, d), 61 (b, e), and 91 (c, f) year calibration windows.

Author contributions. W.H. and S.M jointly conceived the study. W.H. performed the analysis and created all figures. Data interpretation was done by both authors. The manuscript was written by W.H. with input from S.M.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. The authors would like to thank Ryan Batehup for the use of his code (Batehup et al., 2015) which was substantially utilised in this study. W.H. was supported by the Australian Research Council (ARC) via grants FT100100443, DP130104156 and the Potsdam Institute for Climate Impact Research (PIK), member of the Leibniz Association. S.M. was supported by the Australian Research Council (ARC) via grant number FT160100162.

References

Abram, N. J., Mulvaney, R., Vimeux, F., Phipps, S. J., Turner, J., and England, M. H.: Evolution of the Southern Annular Mode during the past millennium, Nature Climate Change, 4, 564–569, https://doi.org/10.1038/nclimate2235, 2014.

Batehup, R., McGregor, S., and Gallant, A. J. E.: The influence of non-stationary teleconnections on palaeoclimate reconstructions of ENSO variance using a pseudoproxy framework, Climate of the Past, 11, 1733–1749, https://doi.org/10.5194/cp-11-1733-2015, 2015.

Bracegirdle, T. J., Holmes, C. R., Hosking, J. S., Marshall, G. J., Osman, M., Patterson, M., and Rackow, T.: Improvements in Circumpolar Southern Hemisphere Extratropical Atmospheric Circulation in CMIP6 Compared to CMIP5, Earth and Space Science, 7, e2019EA001 065, https://doi.org/https://doi.org/10.1029/2019EA001065, 2020.

Cullen, L. E. and Grierson, P. F.: Multi-decadal scale variability in autumn-winter rainfall in south-western Australia since 1655 AD as reconstructed from tree rings of *Callitris Columellaris*, Climate Dynamics, 33, 433–444, https://doi.org/10.1007/s00382-008-0457-8, 2009.

Dätwyler, C., Neukom, R., Abram, N. J., Gallant, A. J. E., Grosjean, M., Jacques-Coper, M., Karoly, D. J., and Villalba, R.: Teleconnection stationarity, variability and trends of the Southern Annular Mode (SAM) during the last millennium, Climate Dynamics, 51, 2321–2339, https://doi.org/10.1007/s00382-017-4015-0, 2018.

Dätwyler, C., Grosjean, M., Steiger, N. J., and Neukom, R.: Teleconnections and relationship between the El Niño–Southern Oscillation (ENSO) and the Southern Annular Mode (SAM) in reconstructions and models over the past millennium, Climate of the Past, 16, 743–756, https://doi.org/10.5194/cp-16-743-2020, https://cp.copernicus.org/articles/16/743/2020/, 2020.

Davey, M., Brookshaw, A., and Ineson, S.: The probability of the impact of ENSO on precipitation and near-surface temperature, Climate Risk Management, 1, 5–24, https://doi.org/10.1016/j.crm.2013.12.002, 2014.

Davis, R. E.: Predictability of Sea Surface Temperature and Sea Level Pressure Anomalies over the North Pacific Ocean, Journal of Physical Oceanography, 6, 249–266, https://doi.org/10.1175/1520-0485(1976)006<0249:POSSTA>2.0.CO;2, 1976.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828, 2011.

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, Journal of Climate, 19, 643–674, https://doi.org/10.1175/JCLI3629.1, 2006.

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, Geophysical Research Letters, 32, L07711, https://doi.org/10.1029/2004GL021236, 2005.

Fogt, R. L., Bromwich, D. H., and Hines, K. M.: Understanding the SAM influence on the South Pacific ENSO teleconnection, Climate Dynamics, 36, 1555–1576, https://doi.org/10.1007/s00382-010-0905-0, 2011.

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, https://doi.org/10.1175/JCLI-D-12-00338.1, 2013.

Gillet, N. P., Kell, T. D., and Jones, P. D.: Regional climate impacts of the Southern Annular Mode, Geophysical Research Letters, 33, L23704, https://doi.org/10.1029/2006GL027721, 2006.

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 Science, 1, 45–79, https://doi.org/10.5194/os-1-45-2005, 2005.

Hauck, J., Völker, C., Wang, T., Hoppema, M., Losch, M., and Wolf-Gladrow, D. A.: Seasonally different carbon flux changes in the Southern Ocean in response to the southern annular mode, Global Biogeochemical Cycles, 27, 1236–1245, https://doi.org/10.1002/2013GB004600, 2013.

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, Journal of Climate, 20, 650–666, https://doi.org/10.1175/JCLI4011.1, 2007.

Hendon, H. H., Thompson, D. W. J., and Wheeler, M. C.: Australian Rainfall and Surface Temperature Variations Associated with the Southern Hemisphere Annular Mode, J. Climate, 20, 2452–2467, https://doi.org/10.1175/JCLI4134.1, 2007.

Huiskamp, W. N. and Meissner, K. J.: Oceanic carbon and water masses during the Mystery Interval: A model-data comparison study, Paleoceanography, 27, PA4206, https://doi.org/10.1029/2012PA002368, 2012.

Huiskamp, W. N., Meissner, K. J., and d'Orgeville, M.: Competition between ocean carbon pumps in simulations with varying Southern Hemisphere westerly wind forcing, Climate Dynamics, pp. 3463–3480, https://doi.org/10.1007/s00382-015-2781-0, 2015.

Jones, J. M., Fogt, R. L., Widmann, M., Marshall, G. J., Jones, P. D., and Visbeck, M.: Historical SAM Variability. Part I: Century-Length Seasonal Reconstructions*, Journal of Climate, 22, 5319–5345, https://doi.org/10.1175/2009JCLI2785.1, 2009.

Karpechko, A. Y., Gillett, N. P., Marshall, G. J., and Screen, J. A.: Climate Impacts of the Southern Annular Mode Simulated by the CMIP3 Models, Journal of Climate, 22, 3751–3768, https://doi.org/10.1175/2009JCLI2788.1, 2009.

Keppler, L. and Landschützer, P.: Regional Wind Variability Modulates the Southern Ocean Carbon Sink, Sci Rep, 9, 7384, https://doi.org/10.1038/s41598-019-43826-y, 2019.

Kwok, R. and Comiso, J. C.: Spatial patterns of variability in Antarctic surface temperature: Connections to the Southern Hemisphere Annular Mode and the Southern Oscillation, Geophysical Research Letters, 29, 1705, https://doi.org/10.1029/2002GL015415, 2002.

Le Quéré, C., Rödenbeck, C., Buitenhuis, E. T., Conway, T. J., Langenfelds, R., Gomez, A., Labuschagne, C., Ramonet, M., Nakazawa, T., Metzl, N., Gillett, N., and Heimann, M.: Saturation of the Southern Ocean CO_2 Sink Due to Recent Climate Change, Science, 316, 1735–1738, https://doi.org/10.1126/science.1136188, 2007.

Lee, S. and Feldstein, S. B.: Detecting Ozone- and Greenhouse Gas–Driven Wind Trends with Observational Data, Science, 339, 563–567, https://doi.org/10.1126/science.1225154, http://www.sciencemag.org/content/339/6119/563.abstract, 2013.

Lenton, A. and Matear, R. J.: Role of the Southern Annular Mode (SAM) in Southern Ocean CO_2 uptake, Global Biogeochemical Cycles, 21, https://doi.org/10.1029/2006GB002714, 2007.

Liu, W., Lu, J., Xie, S.-P., and Fedorov, A.: Southern Ocean Heat Uptake, Redistribution, and Storage in a Warming Climate: The Role of Meridional Overturning Circulation, Journal of Climate, 31, 4727–4743, https://doi.org/10.1175/JCLI-D-17-0761.1, 2018.

Lovenduski, N. S., Gruber, N., Doney, S. C., and Lima, I. D.: Enhanced CO_2 outgassing in the Southern Ocean from a positive phase of the Southern Annular Mode, Global Biogeochemical Cycles, 21, GB2026, https://doi.org/10.1029/2006GB002900, 2007.

Marini, C., Frankignoul, C., and Mignot, J.: Links between the Southern Annular Mode and the Atlantic Meridional Overturning Circulation in a Climate Model, Journal of Climate, 24, 624 – 640, https://doi.org/10.1175/2010JCLI3576.1, 2011.

Marshall, G. J.: Trends in the Southern Annular Mode from Observations and Reanalyses, J. Climate, 16, 4134–4143, https://doi.org/10.1175/1520-0442(2003)016<4134:TITSAM>2.0.CO;2, 2003.

Marshall, G. J. and Bracegirdle, T. J.: An examination of the relationship between the Southern Annular Mode and Antarctic surface air temperatures in the CMIP5 historical runs, Climate Dynamics, 45, 1513–1335, https://doi.org/10.1007/s00382-014-2406-z, 2015.

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, Climate of the Past, 9, 2269–2284, https://doi.org/10.5194/cp-9-2269-2013, 2013.

PAGES 2k Consortium: Continental-scale temperature variability during the past two millennia, Nature Geoscience, 6, 339–346, https://doi.org/10.1038/ngeo1797, 2013.

Previdi, M. and Polvani, L. M.: Climate system response to stratospheric ozone depletion and recovery, Quarterly Journal of the Royal Meteorological Society, 140, 2401–2419, https://doi.org/10.1002/qj.2330, http://dx.doi.org/10.1002/qj.2330, 2014.

Raphael, M. N. and Holland, M. M.: Twentieth century simulation of the southern hemisphere climate in coupled models. Part 1: large scale circulation variability, Climate Dynamics, 26, 217–228, https://doi.org/10.1007/s00382-005-0082-8, 2006.

Russell, J. L., Dixon, K. W., Gnanadesikan, A., Stouffer, R. J., and Toggweiler, J. R.: The Southern Hemisphere Westerlies in a Warming World: Propping Open the Door to the Deep Ocean, Journal of Climate, 19, 6382–6390, https://doi.org/10.1175/JCLI3984.1, 2006.

Silvestri, G. and Vera, C.: Nonstationary Impacts of the Southern Annular Mode on Southern Hemisphere Climate, J. Climate, 22, 6142–6148, https://doi.org/10.1175/2009JCLI3036.1, 2009.

Son, S.-W., Polvani, L. M., Waugh, D. W., Akiyoshi, H., Garcia, R., Kinnison, D., Pawson, S., Rozanov, E., Shepherd, T. G., and Shibata, K.: The Impact of Stratospheric Ozone Recovery on the Southern Hemisphere Westerly Jet, Science, 320, 1486–1489, https://doi.org/10.1126/science.1155939, http://www.sciencemag.org/content/320/5882/1486.abstract, 2008.

Steig, E. J., Mayewski, P. A., Dixon, D. A., Kaspari, S. D., Frey, M. M., Schneider, D. P., Arcone, S. A., Hamilton, G. S., Spikes, V. B., Albert, M., Meese, D., Gow, A. J., Shuman, C. A., White, J. W. C., Sneed, S., Flaherty, J., and Wumkes, M.: High-Resolution Ice Cores from US ITASE (West Antarctica): Development and Validation of Chronologies and Determination of Precision and Accuracy, Annals of Glaciology, 41, 77–84, https://doi.org/10.3189/172756405781813311, 2005.

Sterl, A., van Oldenborgh, G. J., Hazeleger, W., and Burgers, G.: On the robustness of ENSO teleconnections, Climate Dynamics, 29, 469–485, https://doi.org/10.1007/s00382-007-0251-z, 2007.

Thompson, D. W. J. and Solomon, S.: Interpretation of Recent Southern Hemisphere Climate Change, Science, 296, 895–899, https://doi.org/10.1126/science.1069270, 2002.

van Oldenborgh, G. J. and Burgers, G.: Searching for decadal variations in ENSO precipitation teleconnections, Geophysical Research Letters, 32, https://doi.org/10.1029/2005GL023110, 2005.

Villalba, R., Lara, A., Masiokas, M. H., Urrutia, R., Luckman, B. H., Marshall, G. J., Mundo, I. A., Christie, D. A., Cook, E. R., Neukom, R., Allen, K., Fenwick, P., Boninsegna, J. A., Srur, A. M., Morales, M. S., Araneo, D., Palmer, J. G., Cuq, E., Aravena, J. C., Holz, A., and LeQuesne, C.: Unusual Southern Hemisphere tree growth patterns induced by changes in the Southern Annular Mode, Nature Geoscience, 5, 793–798, https://doi.org/10.1038/ngeo1613, 2012.

Visbeck, M.: A Station-Based Southern Annular Mode Index from 1884 to 2005, Journal of Climate, 22, 940–950, https://doi.org/10.1175/2008JCLI2260.1, 2009.

von Storch, H., Zorita, E., and González-Rouco, F.: Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation, International Journal of Earth Sciences, 98, 67–82, https://doi.org/10.1007/s00531-008-0349-5, 2009.

Yun, K.-S. and Timmermann, A.: Decadal Monsoon-ENSO Relationships Reexamined, Geophysical Research Letters, 45, 2014–2021, https://doi.org/10.1002/2017GL076912, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL076912, 2018.

Zhang, Z.-Y., Gong, D.-Y., He, X-Z., L. Y.-N., and Feng, S.-H.: Statistical Reconstruction of the Antarctic Oscillation Index Based on Multiple Proxies, Atmospheric and Oceanic Science Letters, 3:5, 283–287, https://doi.org/10.1080/16742834.2010.11446883, 2010.