Supplementary material for ‘Australasian temperature reconstructions spanning the last millennium’

Joëlle Gergis¹, Raphael Neukom²,³,⁴, Ailie J.E. Gallant¹,⁵, and David J. Karoly¹

¹ School of Earth Sciences, ARC Centre of Excellence for Climate System Science, University of Melbourne, Australia.
² Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, Switzerland.
³ Swiss Federal Research Institute (WSL), Birmensdorf, Switzerland.
⁴ Institute of Geography, University of Zurich, Switzerland
⁵ School of Earth, Atmosphere and Environment, ARC Centre of Excellence for Climate System Science, Monash University, Australia.

Corresponding author

Dr Joëlle Gergis
School of Earth Sciences,
ARC Centre of Excellence for Climate System Science,
University of Melbourne,
VIC 3010, AUSTRALIA
Email: jgergis@unimelb.edu.au
Supplementary material

1. Proxy screening
   1.1 Sensitivity to the proxy screening method: theoretical considerations
   1.1.1 Proxy screening using detrended versus raw data
   1.1.2 Proxy screening using field mean temperatures versus local temperature variations
   1.2 Sensitivity to the screening method: influence on reconstruction results

2. Additional verification
   2.1 Instrumental data experiment
   2.2 Australian temperatures as reconstructed from boreholes
   2.3 Comparison with earlier reconstructions
   2.4 Comparison with noise proxies

3. Assessment of late 20th century temperature variations
   3.1 Underestimation of reconstructed temperatures after 1995

4. Comparison of modelled temperature with reconstructed Australasian temperature
   4.1 GISS-E2-R climate model comparisons

5. References
S1 Proxy screening

Here we describe, discuss and assess the different steps and approaches used to establish the predictor matrix used to develop the Australasian temperature reconstruction. In Appendix A, we describe the proxy screening approach that was used for the reconstruction presented in the main manuscript. In section S1.1.1 (S1.1.2) we discuss the theoretical implications and strengths and weaknesses of using raw versus detrended data (field mean temperatures versus local temperature variations) for proxy screening. The sensitivity of proxy selection to these choices, and the subsequent effect on the results, are then quantitatively assessed in section S1.2.

S1.1 Sensitivity to the proxy screening method: theoretical considerations

The steps involved in the proxy screening process include decisions about whether to detrend the data; adjust the significance of the correlation for high-frequency autocorrelation; and calculate correlations against local grid cells, rather than the field mean.

S1.1.1 Proxy screening using detrended versus raw data

Whether or not to implement detrending has been widely discussed in the literature (von Storch et al., 2004; Burger and Cubasch, 2005; Wahl et al., 2006; Ammann and Wahl, 2007), although mostly in relation to regression model calibration rather than proxy selection. While the influence of using detrended data for model calibration is addressed subsequently in S1.2, we first assess its influence on predictor section.

For data with dominant variations on multiple timescales including a strong, long-term trend, the statistical agreement between a predictor and predictand may be inflated by serial autocorrelation associated with that trend. Detrending the data reduces the autocorrelation associated with the trend, and hence also potentially reduces the strength of the correlation. However, detrending also removes variations that are associated with real geophysical processes (Ammann and Wahl, 2007), which Wahl et al. (2006) argue effectively diminish their importance. For inter-annual temperature data, the removal of long-term trends tends to highlight the variance associated with internal inter-
annual to decadal scale modes such as the El Nino–Southern Oscillation (ENSO).

Figure S1.1 shows that the spatial patterns of the correlations between inter-annual Australasian field mean temperature and local temperatures vary depending on whether or not data are detrended. The large temperature trend in the region over the 20th century influences the correlations computed using raw anomalies. Conversely, the detrended data show a more ENSO-like pattern, the dominant large-scale driver of year-to-year temperature variations in the region (Kiladis and Diaz, 1989; Allan and Lindsay, 1998; Salinger and Mullan, 1999; Jones and Trewin, 2000).

Figure S1.1 The first principal component (PC1) was extracted for the SONDJF HadCRUT3v temperature data. The relative loadings of each grid point are provided for a) PC1 of the raw annual temperature data, b) PC1 for the detrended annual temperature data, c) PC1 of the 11-year smooth raw temperature data, d) PC1 of the 11-year smoothed detrended temperature data.
Figure S1.2 The left panel shows the number of years for which data are available at each grid cell presented in Figure S1.1. The right panel, the associated fraction of missing data calculated as the percentage of the 110-year record. The locations of the 51 proxies in Table A1 are shown as black circles.

The influence of detrending on palaeoclimate site selection was tested using the Australasian HadCRUT3v grid cells that corresponded most closely to each proxy location as predictors. The sensitivity of the resultant PCR reconstruction to site selection was also tested. The correlations between instrumental data at the proxy site and the field mean temperature and local temperatures in a 500 km radius were computed using combinations of raw and detrended data. The presence of any statistically significant correlations identified the sites to be used for the reconstruction. The skill scores for the reconstructions from each combination of site selection methods were then compared. Table S1.1 shows that the weaker skill scores from the detrended data reflect the loss of low-frequency variance associated with the removal of real climate information related to the spatially-coherent trend across the domain, seen in Figure S1.1. The underlying instrumental data length and availability shown in Figure S1.2 may also influence these results.

Table S1.1 Skill scores of PCR reconstructions of Australasian mean temperature computed using different screening methods over the 1921–1990 period. The reconstructions were generated using detrended and raw data for proxy screening. The correlations were computed against Australasian field mean temperatures. For this exercise, the instrumental HadCRUT3v grid cells that corresponded to our proxy locations were used as temperature predictors. The ensemble median RE, and early verification RE skill scores, the root mean square error (RMSE) and the standard deviation of the regression residuals are shown.

<table>
<thead>
<tr>
<th>Proxy screening approach</th>
<th>Ensemble median RE</th>
<th>Early verification RE</th>
<th>RMSE (°C)</th>
<th>Residual SD (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detrended</td>
<td>0.72</td>
<td>0.79</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>Raw</td>
<td>0.77</td>
<td>0.82</td>
<td>0.09</td>
<td>0.10</td>
</tr>
</tbody>
</table>
In summary, the decision to select proxy data using either raw or detrended data influences the composition of the predictor matrix. However, the decision to apply proxy selection also means that the chances of including extraneous noise not associated with temperature variations is substantially reduced by using our method described in Section 2c of the main text. Our results here (and presented later in Section S1.2) show that proxy selection only has a minor influence on resultant reconstructions (e.g. Table S1.4 and Figure S1.4).

**S1.1.2 Proxy screening using field mean temperatures versus local temperature variations**

As part of the proxy screening process, we compared the relationships between each proxy and the field mean and local temperature variations (defined as grid cells within a 500 km radius of the proxy site). The detrended spatial patterns of the correlations in Figure S1.1b illustrate the magnitude of the influence of inter-annual local temperature variations on the field mean. These patterns are inhomogeneous because multiple and different processes influence each location, resulting in different climates (e.g. maritime, continental, tropical, sub-tropical and mid-latitude). Such regimes are defined by characteristic weather systems, the influence of topography, and influences of large-scale climate drivers such as ENSO or the Southern Annular Mode (SAM). These processes result in a ‘noisy’ climate signal at each location, which together with biological or chemical noise associated with palaeoclimate records, means that even if some proxies represent the local temperature variations that ultimately contribute to the Australasian spatial mean temperature variations, these proxies may not display a statistically significant correlation with the full HadCRUT3v field mean.

For example, New Zealand’s topography plays a major role in determining the spatial response of regional temperature patterns to circulation (Salinger and Mullan, 1999). This may explain differences in the temperature response observed between, for example, the North Island and South Island (e.g. Oroko and Pink Pine) tree ring sites.

Figures S1.1c and S1.1d also show that on decadal time scales, coherence in temperature variations across the Australasian domain increases markedly, particularly for the detrended data.
(Figures S1.1d). Thus, we conclude that assessing field-mean variations on the decadal and longer scale is representative of coherent temperature variations across Australasia. The predictor matrices and skill of the resulting reconstructions generated using either the local or field mean approaches are very similar, again highlighting the insensitivity of the results to nuances in the proxy selection (Table S1.4, Figures S1.3–S1.5). For these reasons, local, detrended correlations were used to select our final set of temperature predictors.

**S1.2 Sensitivity to the screening method: influence on reconstruction results**

The sensitivity of the screening method to the spatial inhomogeneity in Australasian mean temperatures and the issue of detrending were tested using the instrumental data described in Section S1.1 above. The following sensitivity analyses are now performed using real proxy data to establish the effect of proxy selection on the temperature reconstruction. The various screening methods are summarized in Table S1.2.

**Table S1.2** Description of the various screening methods used to test the sensitivity of the reconstructions to proxy selection.

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Screening approach</th>
<th>Number of proxies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AR 1 detrending local fieldmean</td>
<td>Significant correlation with at least one grid-cell of the target grid within a search radius of 500 km. All data are linearly detrended before the correlations, AR1 properties are taken into account for the calculations of degrees of freedom</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>No AR 1 no detrending fieldmean</td>
<td>Simple correlation with the reconstruction target time series, no detrending or adjustment of degrees of freedom</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>AR 1 no detrending fieldmean</td>
<td>Same as #2 above but using AR1-corrected degrees of freedom</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>AR 1 detrending fieldmean</td>
<td>Same as #2 above but using detrended data and AR1-corrected degrees of freedom</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>AR 1 no detrending local</td>
<td>Same as #1 above but using non-detrended data</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>No screening</td>
<td>All proxies are used</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>No screening domain proxies</td>
<td>All proxies from only within the reconstruction domain are used</td>
<td>36</td>
</tr>
</tbody>
</table>

These options use various combinations of field mean versus local correlations, detrending versus raw data, and accounting for autoregression in the degrees of freedom for significance
testing (options #1–#5). Option #1 is the approach that was used to develop the final reconstructions presented in the main manuscript and was chosen for the reasons described in Section S1.1. This approach ensures that: i) each predictor is directly connected to regional climate at its location. This excludes extraneous noise but also does not assume that the influence of local climates on the field mean remains stable over time; and ii) potential statistical artefacts that may inflate the proxy-temperature relationship are not included. Option #4 uses the same statistical parameters but correlating the proxies directly with the target field mean. Options #2, #3, and #5 use less rigorous statistical approaches that allow potential geophysical processes leading to a common trend and/or autocorrelation to inflate the proxy-climate relationship. In addition, we present reconstructions applying no screening and i) using all proxies within the wider Australasian domain (option #6), and ii) using only proxies from within the reconstruction domain (option #7). The proxy selections arising from these different screening methods are presented in Table S1.3.
Table S1.3 The proxy networks and respective correlations and significance levels developed using various proxy selection methods. Options #1–#5 only are shown. Pearson correlation coefficients (cor), p-values (p), degrees of freedom (df), the optimal lag selected for each proxy (lag) and an indicator of selection (sel) are provided for each proxy and screening method. NA means that no grid cell within a search radius of 500km had significant (p<0.05) correlations with the proxy. Note that for field mean correlations (last three options), annually resolved coral and ice core records are only selected if their correlation in the target season (SONDJF) is higher than in the winter season (MAMJJA; not shown).
For simplicity, Figures S1.3–S1.5 compare the R28 reconstruction for just the PCR method presented in the main text, with the results based on the range of alternative proxy screening methods. They show that the variations reconstructed for each option are similar. The results always lie within the 2SE uncertainty range of our final reconstruction (option #1), except for a few years for option #4 (Figure S1.3), which only uses nine predictors. This suggests that decadal-scale temperature variability is not highly sensitive to the predictor screening methods. Even a simple average of all available records (including those that contain no significant relationship with temperature during the 20th century) yields surprisingly similar results to our final reconstruction, with some notable centennial-scale variability (blue timeseries in Figure S1.5).

The evolution of the early verification RE skill scores is illustrated in Figures S1.6–S1.8. The different approaches have very similar skill during the early verification period 1900–1930 (Figures S1.6–S1.8). As this early verification period has no significant temperature trend (Mann-Kendall test, α=0.05) that may artificially enhance the verification results, the high early verification RE values indicate skill in the ability of the predictors to capture significant fractions of inter-annual temperature variability. Table S1.4 compares various skill and uncertainty measures of all approaches over the 1896–1990 period, i.e. the period represented by the most replicated proxy nests, where a nest uses all available proxies at time, t.

Another way to assess the influence of using detrended data during the calibration period is to calculate the transfer functions for the reconstruction based on detrended data. We test this by calculating an additional PCR reconstruction where the principal components computed from the proxy data and instrumental target are detrended over the calibration period. That is, in this evaluation exercise the data applied to the multiple linear regressions were all detrended. The regression coefficients were then applied to the raw proxy PCs over the full reconstruction period to obtain the temperature reconstruction. Figures S1.9 and S1.10 show that the difference between raw and detrended calibration are again small, indicating the minor influence of the choice of using raw vs. detrended data on the resultant reconstruction.
Based on the analyses presented in this section, we conclude that the different screening approaches, although representing different statistical and climatological concepts, yield similar reconstruction outcomes. The differences in reconstructed temperatures between the different approaches are mostly smaller than the uncertainties within each approach, arising from calibration errors and ensemble perturbations (as illustrated by the grey shading in Figures S1.3–S1.5).

**Figure S1.3.** 30-year filtered ensemble mean PCR reconstructions using the different methods for proxy selection described in Table 1.3. Option #1 (black line) is the screening method used to generate the reconstruction from the main text with 2SE uncertainties (grey shading; defined as in Neukom et al. (2014) by adding the variance of the ensemble distribution and regression residuals); Option #2 (red dashed line) is the proxy screening based on correlations with the target field mean without corrections; Option #3 (green dotted line) is the same as Option #2 but uses AR1-corrected significance levels; Option #4 (blue dash-dotted line) is the same as Option #2 but using AR1-corrected significance levels and detrended data.

**Figure S1.4.** Same as Figure S1.3 but comparing the PCR reconstruction from the main text (black with grey 2SE-shading; option #1 in Table S1.3) with the reconstruction based on local proxy screening with non-detrended data (red dashed; option #5).
**Figure S1.5.** Same as Figure S1.3 but comparing the PCR reconstruction from the main text (black with grey 2SE-shading; option #1 in Table 1.3) with the reconstruction using all available records (red dashed; option #6) and the reconstructions using all available records from within the reconstruction domain (green dotted, option #7). The blue dash-dotted line represents a simple average of all available records (after scaling each record to mean 0 and standard deviation 1 over the 1921-1990 period and adjusting the sign based on the correlation with the raw/undetrended instrumental target). Note that option #6 (no screening) is comparable to the R28 network, and option #7 (no screening domain proxies only) to R2 network in the early part of the reconstruction.

**Figure S1.6.** The temporal evolution of the early verification (1900-1930) RE skill scores for the ensemble mean of the PCR reconstructions produced using different screening methods. The RE value of each year represents the result of a calibration/verification exercise using only the proxies that are available in this year (nested approach). Black line: the reconstruction from the main text (option #1 in Table 1.3); Red dashed line: Proxy screening based on correlations between the raw target field mean without corrections (option #2); Green dotted line: The same as option #2 but using AR1-corrected significance levels (option #3); Blue dash-dotted line: The same as option #2 but using AR1-corrected significance levels and detrended data (option #4). REs were calculated before adding the regression residual AR(1) noise to the ensemble members, hence the values are not directly comparable to Table 2 in the main text.
Figure S1.7. Same as Figure S1.6 but comparing the early verification RE scores of the reconstruction from the main text (black line, option #1) with the reconstruction based on local proxy screening using raw data (red dashed line, option #5).

Figure S1.8. Same as Figure S1.6 but comparing the early verification RE scores of the reconstruction from the main text (black line, option #1) with the reconstruction using all available records (red dashed; option #6) and the reconstructions using all available records from within the reconstruction domain (green dotted, option #7).
**Figure S1.9.** Influence of using undetrended (black with shaded 2SE envelope) or detrended (red dashed) calibration to develop the R28 PCR temperature reconstruction.

**Figure S1.10.** Influence of using undetrended (black) or detrended (red dashed) calibration on PCR reconstruction skill.
Table S1.4 Skill and uncertainty measures for the different proxy screening methods in a PCR ensemble reconstruction. ‘Early verification RE’: Early (1900-1930) verification RE of the ensemble mean using the most replicated proxy nest. ‘Cor instr vs. recon’: Correlation of instrumental data and ensemble mean reconstruction over 1931-1990; “Cor instr vs. recon (detrended)” same correlation but using linearly detrended data; ‘Residual Std.dev.’: Standard deviation of the verification residuals of all ensemble members; ‘Ensemble Std.dev.’: Average standard deviation among the ensemble members for all years within 1931–1990; ‘RMSE’: Root mean squared error between the instrumental target time series and the ensemble mean reconstruction over 1931–1990. Skill metrics were calculated before adding the regression residual AR(1) noise to the ensemble members, hence the values are not directly comparable to Table 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Early verification RE</th>
<th>Cor instr vs. recon</th>
<th>Cor instr vs. recon (detrended)</th>
<th>Residual Std.dev. (°C)</th>
<th>Ensemble Std.dev. (°C)</th>
<th>RMSE (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR 1 detrending local</td>
<td>0.76</td>
<td>0.73</td>
<td>0.45</td>
<td>0.15</td>
<td>0.07</td>
<td>0.15</td>
</tr>
<tr>
<td>No AR 1 No detrending fldmean</td>
<td>0.77</td>
<td>0.80</td>
<td>0.63</td>
<td>0.13</td>
<td>0.04</td>
<td>0.13</td>
</tr>
<tr>
<td>AR 1 No detrending fldmean</td>
<td>0.75</td>
<td>0.79</td>
<td>0.61</td>
<td>0.13</td>
<td>0.04</td>
<td>0.13</td>
</tr>
<tr>
<td>AR 1 detrending fldmean</td>
<td>0.64</td>
<td>0.83</td>
<td>0.71</td>
<td>0.12</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>AR 1 no detrending local</td>
<td>0.74</td>
<td>0.77</td>
<td>0.57</td>
<td>0.14</td>
<td>0.06</td>
<td>0.14</td>
</tr>
<tr>
<td>No screening</td>
<td>0.74</td>
<td>0.72</td>
<td>0.43</td>
<td>0.15</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>No screening domain proxies</td>
<td>0.66</td>
<td>0.74</td>
<td>0.49</td>
<td>0.14</td>
<td>0.15</td>
<td>0.14</td>
</tr>
</tbody>
</table>

S2. Additional verification

Additional verification of the temperature reconstructions was performed using 1) raw and noise-degraded instrumental data as predictors, 2) Australian land temperatures from published borehole reconstructions (Pollack et al., 2006), 3) Australasian domain temperatures extracted from Mann et al. (2009) and 4) comparison with AR(1) noise proxies (Wahl and Smerdon, 2012; Neukom et al., 2014).

S2.1 Instrumental data experiment

We undertook instrumental verification analyses to test whether we could reasonably reconstruct mean temperature for the whole Australasian field using instrumental data only from grid cells representing the R28 proxy network locations. This analysis is described and discussed in Section 2 of the main text, with further results shown in Figures S2.1–S2.3.
Figure S2.1. Instrumental verification using HadCRUT3v grid points from R28 palaeoclimate record locations as predictors (top). Histogram of mean RE values for 3000-member PCR ensemble (below).
Figure S2.2. Instrumental verification using GHCN stations from R28 palaeoclimate record locations as predictors (top). Histogram of mean RE values for 3000-member PCR ensemble (below).
Figure S2.3. Instrumental experiment based on ten noise reconstructions (top) and combined histogram of mean RE values derived from ten, 3000-member PCR ensembles (below).
S2.2 Australian temperatures as reconstructed from boreholes

Comparison of our Australasian combined land and ocean temperature reconstruction with the Australian land-only borehole-based reconstruction (Pollack et al. 2006) discussed in section 4b is shown in Figure S2.4 below.

**Figure S2.4.** Comparison between Pollack et al.’s (2006) low frequency Australian borehole temperature reconstruction (black) and associated 2SE uncertainties (gray shading) and the 30-year filtered Australasian SONDJF ensemble median temperature reconstructions from AD 1500 onward. Red: PCR reconstruction, green: CPS, blue: LNA, cyan: PaiCo.

S2.3. Comparison with earlier reconstructions

Figure S2.5 compares our reconstruction with the Australasian extraction of the Mann et al. (2008; 2009) global gridded temperature reconstruction. Note that this reconstruction is based on a global predictor set and not only on regional proxies as in our study. The number of records used here exceeds the number of Australasian proxies used in the Mann et al. (2009) reconstruction: only 18 out of the 51 records in Table A1 were available in the Mann et al. (2008) reconstruction (note that all tree ring sites were used as individual predictors in Mann et al. (2008), whereas we use composite records that incorporate many individual sites). While the long-term variations of the two reconstructions generally agree, we identify colder conditions during the first century of the millennium, a warmer 15th century and colder conditions in the mid-19th century.
S2.4. Comparison with noise proxies

Some statistical metrics of skill are sensitive to the choice of the verification period when verification is performed using short time series (e.g. < 30 years). In these cases, a handful of individual years showing a particularly good or poor relationship with the target data can bias the statistical metrics. For this reason, we provide an additional assessment of reconstruction skill by comparing our PCR reconstruction with results from a reconstruction based on AR(1) noise proxies (Wahl and Smerdon, 2012; Neukom et al., 2014).

For each proxy record we create an synthetic proxy time series that has the same length and AR(1) coefficient as the real proxy. We then use these synthetic proxies to develop a 100-member PCR reconstruction. This procedure is repeated 100 times with different synthetic proxies. Figures S2.6-S2.9 compare the RE and CE values of reconstructions computed from the real proxies versus those reconstructions computed from the synthetic proxies for the most replicated proxy nest (representing the period 1900-1990 in R28) and the least replicated proxy nests (representing R2 and the earliest years of the reconstruction in R28).

Note that these RE and CE values are calculated before adding the residual noise (i.e. an estimate of the error between the instrumental values and the reconstruction) to the reconstruction ensemble.
members (see methods, Section 2). This contrasts to Table 2 of the main text, where the skill values were calculated after this perturbation for a better comparison with the other methods. The omission of the residual noise better highlights the skill of the reconstruction, rather than the associated uncertainty.

The real proxies clearly outperform the synthetic proxies in all instances, indicating that the real proxies are capturing at least part of the signal from the instrumental target. The RE and CE values of the real proxies in the R2 network show a bimodal distribution. The peak with lower skill corresponds to the ensemble members where only proxy #2 (Oroko) is included. The members including only proxy #1 (Mt. Read) or both proxies show consistently positive REs and a small fraction (5.1%) of positive CEs even for R2. Note that the verification skill for the ensemble median is considerably higher than the ensemble median of the skill estimates (see Table 2 of the main text).

**Figure S2.6.** Ensemble distribution of RE values of PCR reconstructions for the most replicated proxy nest in R28 based on the real proxies (black) and 100 realizations of AR(1) noise proxies (red).
Figure S2.7. Same as Figure S2.6 but for the CE.

Figure S2.8. Same as Figure S2.6 but for the RE of the R2 network.
S3. Assessment of late 20th century temperature variations

S3.1 Underestimation of reconstructed temperatures after 1995

In the years 1996–2001 reconstructed temperatures are mostly lower than instrumental data (Figure 2, main text). However, this difference is not caused by the proxy records’ inability to register exceptionally warm temperatures (a phenomenon called ‘divergence’ in the literature, see e.g. Esper and Frank, 2009) for the following reasons:

1. ‘Unequal attention’: Esper and Frank (2009) provide a number of possibilities to incorrectly detect a divergence problem. Our case is an example of ‘unequal attention’ (see their Figure 1i), where differences between reconstructed and instrumental temperatures of similar magnitude occur also in other years during the overlap period. In our case, there is also an underestimation of instrumental temperatures around 1940, a rather cool phase. Hence, the difference at the end of the reconstruction is very likely to be a ‘normal’ calibration issue unrelated to the particularly warm instrumental temperatures over the 1996–2001 period.
2. **Sub-regional temperature variations**: From 1996–2001 the number of proxies declines (between 13 in 1996 and 7 in 2001, see Table 1 of the main text). The most important candidates for a divergence problem are the two tree ring records Mt Read and Oroko, which both cover the full reconstruction period. Figure S3.1 shows a variation of the PCR reconstruction using these two proxies only (red curve). An alternative PCR reconstruction is provided using the instrumental data that correspond to these two tree ring records (Cook *et al.* 2006) as predictors (green curve). The apparent discrepancy remains practically unchanged if instrumental predictors are used. Hence, the difference between our reconstruction and the instrumental target is likely to stem from differences between sub-regional temperatures (from New Zealand and Tasmania) and the Australasian temperature mean during this period. The relationship between the tree ring data and local temperatures remains robust during this period.

The above factors show that divergence is unlikely to explain the post-1995 discrepancies in our results. All proxy-based reconstructions are imperfect and local temperatures may not always be representative of large-scale fluctuations, particularly from a small network of proxies. As previously discussed, this is particularly evident in the lower replication of proxies in the R3 and R2 networks. The analyses presented in Figures 3 and 4 of the main manuscript and Appendix C illustrate that this reduced replication leads to less robust reconstructions with increased uncertainties.
**Figure S3.1.** Black: Instrumental target data 1900-2001. Red: PCR Reconstruction ensemble median using only proxies 1 and 2 (see Table 1) as predictors. Green: Same as red but using the corresponding instrumental temperature station data as proxies (see Cook et al. (2006) for data description).

**S4. Comparison of modelled temperature with reconstructed Australasian temperature**

**S4.1. GISS-E2-R climate model comparisons**

Figure S4.1 shows an alternative to Figure 5 presented in the main manuscript using a preindustrial 1500–1850 base period instead of anomalies relative to 1961–1990 levels.

**Figure S4.1.** A comparison of the 30-year filtered Australasian SONDJF ensemble median temperature reconstruction for PCR (red), CPS (green), LNA (blue) and PaiCo (cyan) mean of the three members of GISS-E2-R model ensemble (solid black line). The three-member ensemble range for the model data is denoted by grey shading. All anomalies are calculated relative to a preindustrial A.D. 1500–1850 base period, contrasting to Figure 5 that shows anomalies relative to a 1961–1990 base period.
**Figure S4.2.** 30-year filtered simulations of Australasian mean SONDJF temperature for the period AD 1001–2000 in the GISS E2-R model. Three ensemble members with full forcing (blue; bold blue is the ensemble mean), a historical run with natural forcing only (green: spliced to the past1000 run) and 1000 years from a pre-industrial control simulation (red; AD 800 forcing conditions). Anomalies are shown relative to the 1000-1850 period (full and natural forcing) and to the first 850 years of the segment shown for the control simulation.

**Figure S4.3.** The distribution of the changes in Australasian mean SONDJF temperature between consecutive 50-year periods of a 1200-year GISS-E2-R pre-industrial control simulation.
S5. References


