Supplement of

Continental-scale temperature variability in PMIP3 simulations and PAGES 2k regional temperature reconstructions over the past millennium

Correspondence to: H. Goosse

The copyright of individual parts of the supplement might differ from the CC-BY 3.0 licence.
Sect. S1: processing of the model time series

Model data were processed to correspond to the temporal and area-weighted spatial averaging of the PAGES 2k temperature reconstructions (PAGES 2k Consortium, 2013) for each region as follows:

1. Antarctica (annual): 90° S-60° S; 180° W-180° E
2. Arctic (annual): 60° N-90° N; 180° W-180° E
3. Asia (June-August): 23.5° N-55° N; 60° E-160° E
4. Australasia (September-February): 50° S-0° S; 110° E-180° E
5. Europe (June-August): 35° N:70° N ; 10° W-40° E
6. North America (annual)*: 30° N:55° N ; 130° W-75° W
7. South America (December-February): 65° S-20° S ; 75° W-30° W

Note that all simulations for North America were bilinearly interpolated to the HadCRUT3V latitude-longitude grid before the grid box centred at 52.5° N; 77.5° W was removed from the averaging calculation to match processing of the instrumental predictand (PAGES 2k Consortium, 2013).

The ocean was masked for Antarctica, Asia, Europe and South America regions using the respective binary (1, 0) land masks for each simulation. No significant difference was found between the use of fractional land masking (proportion between and including 0 and 1) and the binary mask common to all simulations for the land-only PAGES 2k reconstruction equivalent.

As the GISS-E2-R control simulation is known to contain a drift from non-equilibrated initial conditions (Schmidt et al., 2014), the transient as well as the control simulation from that model have been detrended by subtracting a low-frequency loess fit that has been estimated from the corresponding time period of the control simulation from each time series.

Sect. S2: specific implementation of some methods

S2.1 Probabilistic and climatological consistency

The two concepts of probabilistic and climatological consistency can be seen as two alternative ways to evaluate biases and spreads in simulations and reconstructions. In the current study, however, the analyses use temperature anomalies from long-term averages, and hence the bias is always zero by construction. Thus, our analysis mainly assesses two different aspects of spread in the distributions of the regional temperature reconstructions and climate model simulations.

The quantile-quantile (r-q-q) plots displayed to analyse the climatological consistency show the difference between the simulated and the target quantiles. Residuals should approach zero for a consistent simulation (a flat line in the plots). Offsets relative to y=0 on the quantile-quantile plot indicate biases between the simulation and the target. Slopes in the residuals indicate underestimation or overestimation of the variance, i.e. excessively narrow or wide distributions. We refer to such cases as being under- or over-dispersive. Negative slopes occur if the simulated variance is smaller than that of the target and positive ones for larger simulated variance. If a simulation and a reconstruction are consistent, the difference in their quantiles should be close to zero. Consequently, the plot should be approximately flat. If for low values of the reconstruction the residual quantile simulated minus reconstruction are always negative and always positive for positive value of the reconstructions (positive slope),
the simulated ensemble is too broad and thus the simulations have larger simulated variance compared to the reconstruction.

To assess probabilistic consistency, we test whether the occurrence frequencies of the simulation ensemble agree with those of the verification target, within limits of uncertainty. At each time step, we identify the rank of the temperature reconstructions within the set formed by the combination of the simulation ensemble and those temperature reconstructions (Anderson, 1996). Flatness of the histograms is thus a necessary condition for our simulation ensemble to be considered as a reliable representation of the target. The histograms visually highlight biases (meaning here an offset in mean between the target and the ensemble) and differences in ensemble variance. Over-dispersion (ensembles that are too wide) and under-dispersion (ensembles that are too narrow) are identified by dome- or U-shaped histograms, respectively. Such shapes imply that the target data are too often close to the central rank or too often on the outer ranks (i.e., far from the mean of the ensemble of simulations). Slopes in the histograms reveal biases, with positive (negative) slopes suggesting the target data are ranked high (low) too often.

S2.2. Superposed epoch analysis

The response to volcanic aerosol forcing is evaluated at interannual and multidecadal time scales for two different external forcing estimates (Gao et al., 2008; Crowley and Unterman, 2013) that have been used as last-millennium boundary conditions in the PMIP3-CMIP5 simulations (Schmidt et al., 2011, 2012).

The volcanic composite at interannual timescales is generated by first selecting the 12 strongest volcanic events. The mean from 5 years before to 10 years after the date of the peak eruption is then computed for the forcing sequence as well as for the simulated and reconstructed temperature sequences. In the case of the multidecadal composites, the time series are first filtered with a 40-year low pass filter using least-squares coefficients (Bloomfield, 1976). For the multidecadal composites, the 5 strongest events are selected and the means from 40 years before to 40 years after the eruption are calculated, following Masson-Delmotte et al. (2013). All the events are individually selected for each of the PAGES 2k regions making use of the latitudinal discretization of the volcanic forcing.

The composites for the strongest multidecadal changes in the solar forcing are based on low solar forcing periods selected to be the same as in Fig. 5.8 of the IPCC AR5 (Masson-Delmotte et al., 2013) for the sake of a better comparison. This corresponds to seven 80-year time windows centred on the years 1044, 1177, 1451, 1539, 1673, 1801 and 1905.

S2.3 Framework for evaluation of climate model simulations: $U_R$ and $U_T$ statistics

The statistical model underlying the framework developed by Sundberg et al. (2012, henceforth SUN12), Hind et al. (2012) and Moberg et al. (2015) has similar components to the one used in detection and attribution studies (see section 5.3), but there are some differences. An important similarity is the idea that temperature variations can be expressed as a sum of forced and unforced variability. The two frameworks explicitly distinguish internal variability in simulations and in observations, which can consist of instrumental observations or, as in this study, proxy-based climate reconstructions.

The SUN12 framework also explicitly accounts for error variance in the observations, such as non-climatic noise in proxy data. It even allows this type of error to vary with time, if
such information is available. Despite similarities in the underlying assumptions, the main purposes of the SUN12 and detection and attribution approaches differ. While detection and attribution studies seek to identify the forced response in observations, the SUN12 framework was developed as a tool for evaluating forced simulations, with the aim of testing if one simulation significantly fits observations better than another simulation or to rank a set of plausible simulations. In the current study, this framework is mainly used to investigate the common behaviour of all simulations by means of how well they agree with the different regional reconstructions.

$U_R$ and $U_T$ are calculated here for each forced simulation, using PAGES 2k regional temperature reconstructions as the observational basis and a time resolution of non-overlapping 15-year averages. Three types of calculations have been done: separately for each region, combining information from all seven regions and combining regions only within each hemisphere, using equal regional weights (see Moberg et al., 2015). Whenever a certain control simulation is not sufficiently long, its data sequence is extended by repetition and concatenation. For the COSMOS ensembles with high and low solar forcing, and for the GISS ensemble, metrics are calculated for each individual simulation and for the entire ensembles, following Moberg et al. (2015).

The statistical framework by Sundberg et al. (2012) requires that all proxy-based temperature reconstruction time series are re-calibrated against instrumental records to suit certain assumptions. Therefore, such a re-calibration was done here, but note that this is specific for the calculation of $U_R$ and $U_T$ statistics and is not applied for any other diagnostics.

To obtain appropriate calibration target data series, gridded instrumental temperature data were used and averaged over exactly the same regions and seasons as explained in Sect. S1 for the models. To comply with the different boundaries and land/sea masks used, the respective instrumental series were derived from CRUTEM4 (Jones et al., 2012) for regions 5 and 6, HadCRUT4 (Morice et al., 2012) for regions 2, 4 and 6 and CRU TS3 (Harris et al., 2014; as updated and available on the KNMI Climate Explorer, http://climexp.knmi.nl/, on April 23, 2014) for region 3. For region 1, we used the same instrumental target series as the PAGES 2k Consortium (2013). Re-calibration was made for the same calibration periods as used by the PAGES 2k Consortium. Each instrumental target series was arbitrarily assumed to contain 10% noise variance. Sensitivity experiments were also made with assumptions of 5% and 15% noise. This had no effect on any main conclusions.

For simplicity, it is assumed here that each proxy record has the same statistical precision over its entire length, despite the fact that their precision typically decreases back in time as the number of contributing local proxy series decreases. Therefore, the derived measures are only approximate values, but a more accurate treatment would require detailed work far beyond the scope of this study. As for some other methods in this study, the $U_R$ and $U_T$ analysis uses anomalies from long-term averages to avoid systematic climatological bias influencing the results.

S2.4 Detection and attribution

Detection and attribution techniques provide an estimate of the magnitude of the forced response in a reconstruction, with an uncertainty estimate. These techniques can be used to determine the relative contribution by different forcings simultaneously to a period or climatic event, with uncertainty estimates reflecting if the contribution of different forcings
can be separated from each other and from climate variability (see Bindoff et al., 2013; Hegerl and Zwiers, 2011). To estimate the different contributions from several individual forcings, it is necessary to have access to separately forced simulations with each individual forcing. This was not possible in this study however, because we are only using models driven by all forcings together; hence we focus here on estimating the magnitude of the overall forced response.

Detection and attribution studies rely on a multiple regression of reconstructions onto the response expected by different individual contributing forcings. This assumes that climate models approximately capture the response to individual forcings in shape (e.g., pattern in time or spatial pattern of the response), but may misrepresent the magnitude of the overall response. This is a reasonable assumption since the magnitude of the response to forcings is affected by uncertainty in the transient climate sensitivity. Moreover, the magnitude of forcings itself is also often uncertain, such as for the low-frequency component of solar forcing (see e.g., Schmidt et al., 2011, 2012). A difficulty in the application of detection and attribution methods to the last millennium is accounting for uncertainty in both reconstructions and forcings. This can be addressed to some extent by using multiple reconstructions and forcing estimates (e.g., Schurer et al., 2014), but a more systematic approach is desirable.

The detection and attribution framework applied here has been extensively used for instrumental data (Bindoff et al., 2013) and to some extent for paleoclimatic reconstructions (see Hegerl et al., 2007; Schurer et al., 2014). This approach calculates a possible scaling range for the response to the external forcing in the reconstruction (equation S1) based on total least squares regression (Allen and Stott, 2003):

\[
Y(t) = \sum_{i=1}^{m} \beta_i (X_i(t) - \gamma_i(t)) + \gamma_0(t) \tag{S1}
\]

where \(Y\), the reconstructed temperature, is equal to a linear combination of \(m\) different model fingerprints \(X_i\) (where \(m\) in this analysis is always equal to 1 as only the response to all the forcings together is analysed here) multiplied by a scaling factor \(\beta_i\). Each model simulation has associated internal variability \(\gamma_i\) and the reconstructions contain a realization of internal variability \(\gamma_0\). The scaling factors \(\beta_i\) determine the amplitude of the fingerprints in the reconstructions. A range of scaling factors is obtained using samples of internal variability taken from model simulations. A forcing is said to be detected if a scaling value of zero is rejected at some significance level, for example, the 5% level. To evaluate the self-consistency of the regression results, the residual of the fit is checked against estimates of model-based internal variability. This is the same method as used in Schurer et al. (2013).

**Sect. S3: correlation between simulated and reconstructed time series.**

Figure S1 displays the correlation between the 23-year Hamming filtered model simulation results and temperature reconstructions for individual regions. This illustrates the agreement between the contribution of radiative forcing on observed temperatures and in the model simulations. The highest correlation values are obtained for the Arctic region in most simulations (Figure S1a). Correlations for the North American pollen-based reconstruction and for Australasia and Europe tend to be highly significant. Correlations tend to be non-
significant for the North America tree reconstruction, and for the South American and Antarctic reconstructions.

If we consider the available single-model ensembles (Figure S1b; COSMOS and GISS), the correlation of the ensemble mean with the regional temperature reconstructions is always higher than the average of all individual member correlations. The ensemble averaging reduces the internal variability present in the simulated series in favour of the response to the external forcing common in simulation results and reconstructed temperature.

Sect. S4: EOF analysis for GISS and COSMOS ensembles.

Investigating the variability in the GISS and the COSMOS ensemble simulations provides insights into the intra-model spread. The (detrended) GISS simulations show a very coherent picture with similar loadings and variance explained by the leading EOF (~80-90 %) for the different regions within the single ensemble members (Fig. S7). However, the COSMOS simulations have a larger spread of the variance explained by the leading EOF. The larger heterogeneity in the COSMOS simulations might be indicative of a larger amount of internal variability and hence less externally forced spatial coherence among the regions. For the ensemble with the larger scaling of the solar forcing (COSMOS high, Fig. S7) the amount of variance represented by the leading EOF is larger compared to the weaker scaling (COSMOS low, Fig. S7), indicating a larger common forced signal in the different ensemble members.

Sect. S5: correlation between hemispheres.

An analysis of coherence between hemispheric temperatures, calculated herein simply by weighting the individual regions according to their area (Fig. S10), confirms the results of Neukom et al. (2014) in the sense that the two hemispheres are significantly correlated during most of the last millennium in model simulations. The control simulations indicate a natural tendency for inter-hemispheric correlation in models. Nevertheless, the correlation during periods with strong external forcing clearly exceeds the range derived from control simulations (not shown). In the reconstructions, the Southern Hemisphere experienced temperature anomalies opposite to the ones in the Northern Hemisphere during long periods of the first half of the millennium, indicative of non-coherence between the two hemispheres and potentially unforced variability.


The simulated response to the Gao et al. (2008) forcing has an amplitude of the order of -1 to -0.5 °C in all regions (Fig. S11). This means that the simulated response is, as for the Crowley and Untermann (2012) forcing (Fig. 8), larger than the reconstructed one, in particular for Australia and Antarctica.
Additional references


Supplementary table S1: additional information on model simulation sources

<table>
<thead>
<tr>
<th>Model</th>
<th>past1000</th>
<th>historical extension</th>
<th>Source past1000</th>
<th>Source historical extension</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCSM4</td>
<td>r1i1p1</td>
<td>r1i2p2</td>
<td>PMIP3(^1)</td>
<td>PMIP3</td>
<td>Landrum et al. (2013)</td>
</tr>
<tr>
<td>CESM1</td>
<td>Single continuous simulation (not contained in the CMIP5/PMIP3 database)</td>
<td>Flavio Lehner (<a href="mailto:lehner@climate.unibe.ch">lehner@climate.unibe.ch</a>)</td>
<td>Flavio Lehner (<a href="mailto:lehner@climate.unibe.ch">lehner@climate.unibe.ch</a>)</td>
<td>Lehner et al. (submitted.)</td>
<td></td>
</tr>
<tr>
<td>CSIRO-Mk3L-1-2</td>
<td>r1i1p1</td>
<td>r1i1p1</td>
<td>PMIP3</td>
<td>PMIP3</td>
<td>Phipps et al. (2013)</td>
</tr>
<tr>
<td>GISS-E2-R</td>
<td>r1i1p12[1,4,7]</td>
<td>r1i1p12[1,4,7]</td>
<td>PMIP3</td>
<td>PMIP3</td>
<td>Schmidt et al. (2013)</td>
</tr>
<tr>
<td>HadCM3</td>
<td>r1i1p1</td>
<td>exists, but is not in CMIP5/PMIP3 database</td>
<td>PMIP3(^2)</td>
<td>Andrew Schurer (<a href="mailto:aschurer@staffmail.ed.ac.uk">aschurer@staffmail.ed.ac.uk</a>)</td>
<td>Schurer et al. (2013)</td>
</tr>
<tr>
<td>IPSL-CM5A-LR</td>
<td>r1i1p1</td>
<td>exists, but is not in CMIP5/PMIP3 database</td>
<td>PMIP3</td>
<td>PMIP3</td>
<td>Dufresne et al. (2013)</td>
</tr>
<tr>
<td>MPI-ESM-P</td>
<td>r1i1p1</td>
<td>r1i1p1</td>
<td>PMIP3</td>
<td>PMIP3</td>
<td>Jungclaus et al. (2014)</td>
</tr>
<tr>
<td>ECHAM5/MPIOM</td>
<td>continuous simulations (pre-PMIP3, no rip code)</td>
<td>CERA(^3)</td>
<td>CERA</td>
<td>Jungclaus et al. (2010)</td>
<td></td>
</tr>
</tbody>
</table>


\(^2\) [http://badc.nerc.ac.uk/browse/badc/euroclim500/data/ALL/r1](http://badc.nerc.ac.uk/browse/badc/euroclim500/data/ALL/r1)

\(^3\) [http://cera-www.dkrz.de/CERA](http://cera-www.dkrz.de/CERA)
Figure S1. a) Correlations between 23-year Hamming filtered PAGES2k temperature reconstructions and climate model simulations. Dots represent the correlation between each regional reconstruction (see legend for region-colour) and the simulation averaged over the corresponding domain. Filled (unfilled) circles stand for significant (non-significant) correlation values. b) As in a), but focused on the models with ensembles of simulations; individual ensemble member correlations are shown in grey and the ensemble average in colour.
Figure S2. Normalized spectra of pre-PMIP3 (dashed) and PMIP3 (solid) simulations (grey) and reconstructions (red) for six PAGES 2k regional reconstructions for the period 850 to 2000 CE. The spectra were computed from the normalized reconstructed and simulated regional temperatures using a 100 years Tukey-Hanning filter (Priestley, 1982). The simulations using solar forcing with higher (lower) variability are also highlighted in dark (light) grey.
Figure S3. Climatological consistency: residual quantile-quantile plots for the full period for all the regions. In the left column, the uncertainty is neglected in the computations, in the middle column the original uncertainty divided by a factor $\sqrt{15}$ is used to take into account the smoothing while the original uncertainty is applied for the right column. There is no middle column for North American reconstructions because of their resolution. Positive and negative slopes or large differences from 0 emphasize lack of consistency.
Figure S4. Probabilistic consistency for all the regions. The $\chi^2$ goodness-of-fit statistic is applied to evaluate the consistency between observed rank count and the flat null hypothesis. The statistic can be decomposed to test for individual deviations like bias or spread (Jolliffe and Primo, 2008), as in Bothe et al. (2013a, b). In the left column, the uncertainty is neglected in the computations, in the middle column the original uncertainty divided by a factor $\sqrt{15}$ is used to take into account the smoothing while the original uncertainty is applied for the right column. There is no middle column for North American reconstructions because of their resolution. U- or dome-shaped features highlight lack of consistency.
Figure S5. Skill metric for the individual models for all periods (bars from left to right: 850-1350, 1350-1850, 850-1850, 850-2000 CE). In the left column, the uncertainty is neglected in the computations, in the middle column the original uncertainty divided by a factor $\sqrt{15}$ is used to take into account the smoothing while the original uncertainty is applied for the right column. There is no middle column for North American reconstructions because of their resolution. When the skill is undefined no bar is shown. Positive values indicate skill in this simple evaluation.
a) Correlation for the entire period

b) Correlation for the preindustrial period

Figure S6. Correlations among the PAGES 2k regions for the different models using detrended time series filtered with a 23-year Hamming filter. a) full period 1012 CE – 1978 CE, b) pre-industrial period 1012 CE -1850 CE. The upper left triangle represents the correlations for the forced simulations while the lower right triangle represents the correlations for the control runs (based on the full length of the control runs).
Figure S7. Same as Figure 6a for the leading EOF of the COSMOS ensemble with low (a) and high (b) solar activity changes and the GISS ensemble (c) models over the period 850–2004 AD. The figure shows the spread among the single members for those models with multiple realizations. The eigenvectors are based on the covariance matrix with respect to temperature anomalies for the period 850–1850. Values in parentheses relate to the amount of variance represented by the leading EOF. The time series were filtered with a 23-year Hamming filter and were linearly detrended afterwards. Within their specific experimental setup (COSMOS low, COSMOS high, GISS) the individual simulation members show similarities related to both the amplitude of the temperature anomalies and the variance represented by the leading EOF.
Figure S8. Leading EOFs of the near-surface temperature simulated by each CMIP5/PMIP3 model and in reconstructions over the period 850–1850 CE. The time series were filtered with a 23-year Hamming filter and were linearly detrended before the covariance matrix was calculated. Values in parentheses relate to the amount of variance represented by the leading EOF. The difference to Fig. 6a in the main text relates to a different basis for the calculation of the EOFs corresponding to the pre-industrial period. The pre-industrial EOF pattern is similar to Fig. 6a, albeit with differences in the amplitude of temperature anomalies in individual regions and the amount of variance represented by the leading EOF.
Figure S9. 100-year moving Tukey window correlations between all PAGES 2k regions for the PAGES 2K reconstructions (blue) and PMIP3 models (8 models in orange, multi-model mean in red) and observations from HadCRUT4 (black). Each 100-year segment is linearly detrended beforehand. Grey shading illustrates not significant correlation at the 5% level.
Figure S10. 100-year moving Tukey window correlations between hemispheric averages for the PAGES 2k reconstructions (blue) and PMIP3 models (8 models in orange, multi-model mean in red) and observations from HadCRUT4 (black). Each 100-year segment is linearly detrended beforehand. Grey shading illustrates not significant correlation at the 5% level.
Figure S11. Same as Figure 8 but for the events selected in the Gao et al. (2008) reconstruction.

Figure S12. Superposed Epoch Analysis of the impact of the volcanic activity at multidecadal timescales in the reconstructed and simulated temperatures. Superposed composites of temperature responses during time intervals in which the years with peak negative forcing in the Crowley and Unterman (2012) volcanic reconstruction are aligned. The composite is produced by selecting the 5 strongest volcanic events, and a composite of the 30-year low pass filtered temperature series from 40 years before to 40 years after the date of the peak eruption. All the other elements are the same as in Figure 8.
Figure S13. Same as Figure S12 but for the Gao et al. (2008) forcing.

Figure S14. Superposed Epoch Analysis of the impact of the solar activity at multidecadal timescales in the reconstructed and simulated temperatures. Superposed composites of the temperature response during selected periods in which the solar forcing was lowest were performed (see text for details). Panels show results for reconstructions in six PAGES 2k regions and for model experiments performed using the volcanic forcing by: (top) the Crowley and Untermann (2012); and (bottom) Gao et al. (2008). Each panel indicates the reconstructed (dashed lines) and simulated (solid) composites of the temperature response for the same events for two different regions (see each panel for legend). The colour shading indicates the complete range of simulated temperature responses.
Figure S15: Correlation ($U_R$) and distance ($U_T$) statistics for PAGES 2k regions, with hemispheric and global combinations of all regional data, in the period 856–1350 CE. Positive $U_R$ indicates that simulations and reconstructions have a positive correlation and that they share an effect of temporal changes in external forcings. Negative $U_T$ indicates that a forced simulation is closer to the observed temperature variations than its own control simulation. Coloured dots: individual simulations. Diamonds: ensemble-mean results for COSMOS and GISS models. Dashed lines show one-sided 5% and 1% significance levels. Note the reversed vertical axis in the $U_T$ graphs.
Figure S16: Same as Figure S15 but for the period 1356–1850.
Figure S17: Same as Figure S15 but for the period 861–2000.
Figure S18: Distance ($U_T$) statistics computed for a direct comparison between the high vs. low solar COSMOS simulation ensembles, using the method of Moberg et al. (2015, Appendix B4), for PAGES 2k regions and four different analysis periods. A negative $U_T$ (upwards in the graph) indicates that the high solar simulation ensemble is closer to the observed temperature variations than the low solar ensemble. Dashed lines show two-sided 5% significance levels for the null hypothesis that the two simulations are equivalent. Results for each region are indicated with their abbreviated names. Results where regions are combined are shown with blue symbols: All regions (circle), Northern Hemisphere regions (diamond), Southern Hemisphere regions (cross).