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

PAGES2k-PMIP3 group: Oliver Bothe[1], Michael Evans[2], Laura Fernández Donado[3], Elena Garcia Bustamante[4], Joelle Gergis[5], J. Fidel Gonzalez-Rouco[3], Hugues Goosse[6,a], Gabi Hegerl[7], Alistair Hind[8,9], Johann Jungclaus[1], Darrell Kaufman[10], Flavio Lehner[11,12], Nicholas McKay[10], Anders Moberg[8], Christoph C. Raible[12], Andrew Schurer[7], Feng Shi[13], Jason E. Smerdon[14], Lucien von Gunten[15], Sebastian Wagner[16], Elliott Warren[17], Martin Widmann[17], Pascal Yiou[18] Eduardo Zorita[16].

1. Max Planck Institute for Meteorology, Hamburg, Germany
2. Department of Geology & ESSIC, University of Maryland, USA
3. Institute of Geoscience (UCM-CSIC), Dpt. Astrophysics and Atmospheric Sciences, University Complutense Madrid, Spain
4. Department of Physics. University of Murcia, Spain.
5. School of Earth Sciences, University of Melbourne, Australia
6. ELIC/TECLIM, Université Catholique de Louvain, Belgium
7. School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom
8. Department of Physical Geography, Stockholm University, Sweden
9. Department of Mathematics, Stockholm University, Sweden
10. School of Earth Sciences & Environmental Sustainability Northern Arizona University
11. National Center for Atmospheric Research, Boulder, USA
12. Climate and Environmental Physics and Oeschger Centre for Climate Change Research, University of Bern, Switzerland
13 Key Laboratory of Cenozoic Geology and Environment, Institute of Geology and Geophysics, Chinese Academy of Sciences, 100029 Βειφινγ, Χηινα.
14. Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York, USA
15. PAGES International Project Office, Falkenplatz 16, 3012 Bern, Switzerland
16. Institut for Coastal Research, Helmholtz Zentrum Geesthacht, Germany
17. School of Geography, Earth and Environmental Sciences, University of Birmingham, United Kingdom
18. Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, CE Saclay l'Orme des Merisiers, 91191 Gif-sur-Yvette, France

[a] corresponding author

**Final Version: 27-11-2015**

## Abstract

Estimated external radiative forcings, model results and proxy-based climate reconstructions have been used over the past several decades to improve our understanding of the mechanisms underlying observed climate variability and change over the past millennium. Here, the recent set of temperature reconstructions at the continental-scale generated by the PAGES 2k project and a collection of state-of-the-art model simulations driven by realistic external forcings are jointly analysed. The first aim is to estimate the consistency between model results and reconstructions for each continental-scale region over the time and frequency domains. Secondly, the links between regions are investigated to determine whether reconstructed global-scale covariability patterns are similar to those identified in model simulations. The third aim is to assess the role of external forcings in the observed temperature variations. From a large set of analyses, we conclude that models are in relatively good agreement with temperature reconstructions for Northern Hemisphere regions, particularly in the Arctic. This is likely due to the relatively large amplitude of the externally forced response across northern and high latitude regions, which results in a clearly detectable signature in both reconstructions and simulations. Conversely, models disagree strongly with the reconstructions in the Southern Hemisphere. Furthermore, the simulations are more regionally coherent than the reconstructions perhaps due to an underestimation of the magnitude of internal variability in models or to an overestimation of the response to the external forcing in the Southern Hemisphere. Part of the disagreement might also reflect large uncertainties in the reconstructions, specifically in some Southern Hemisphere regions, which are based on fewer paleoclimate records than in the Northern Hemisphere.

## 1. Introduction

The past millennium is an important period for testing our understanding of the mechanisms that give rise to climate system variability (e.g., Masson-Delmotte et al., 2013). Constraints on, and uncertainties in, external radiative forcings that drive climate change have been extensively documented (e.g., Schmidt et al., 2011, 2012). Such radiative forcing data sets can be used to drive climate simulations using the same model versions that are applied to simulate future climate changes. This allows an evaluation of the relative importance of the various forcings over time, while comparisons of past and future climate simulations place $20^{th}$-century climate variability within a longer context (e.g., Schmidt et al., 2014; Cook et al. 2015). Additionally, the availability of high quality paleoclimatic observations for the last 1000 years permits the reconstruction of regional, hemispheric and global scale climate variability (e.g., Mann et al., 1999; Cook et al., 1999, 2004, 2010; Mann et al., 2009; Jones et al., 2009; PAGES 2k Consortium, 2013, 2015; Masson-Delmotte et al., 2013; Neukom et al., 2014). As a result, the past millennium has become a useful test case for evaluating climate and earth system models used within the Intergovernmental Panel on Climate Change (IPCC) fifth assessment report (Flato et al., 2013; Bindoff et al., 2013).

Paleoclimate reconstructions provides opportunities to test the fidelity of modelled processes and their role in explaining past climatic variations. Reconstructions and simulations can also be used jointly to evaluate estimates of climate sensitivity to external radiative forcing (e.g., Hegerl et al., 2006; Braconnot et al., 2012; Masson-Delmotte et al., 2013). Comparisons across many realizations of simulated climate are used to assess the extent to which characteristic climate statistics are accurately simulated, as well as to disentangle unforced and forced patterns (e.g., Hargreaves et al., 2013; Bothe et al., 2013ab; Neukom et al., 2014; Coats et al., 2015a,b). Estimates of the unforced variability of the climate system may be made from unforced simulations, or from the residual obtained when the forced signal is removed from climate reconstructions, using realistically forced model experiments (Schurer et al., 2013).

Furthermore, simulations can provide the basis for the design of observing network arrays (Comboul et al., 2015). Simulation results also provide a test bed for paleoclimatic reconstruction algorithms within so-called pseudo-proxy experiments (e.g., Zorita et al., 2003; Hegerl et al., 2007; Smerdon, 2012; Lehner et al., 2012; Tingley et al., 2012; Wang et al., 2014; Smerdon et al., 2015). All of these purposes, which are also pursued within the historical period and with comparison to direct climate observations (Bindoff et al., 2013; Ding et al., 2014), are potentially extended by the longer time interval made possible by analysis over the past millennium.

However, obtaining unequivocal conclusions from the comparison between reconstructions and simulation results over the past millennium remains difficult due to uncertainties in climate and forcing reconstructions, the simplified world represented by climate models, and the relatively weak forced signal in the pre-industrial part of the past millennium compared to internal climate variability (e.g., Moberg, 2013). Reconstructions and simulations are two different representations of the behaviour of the actual climate system, and this creates multiple uncertainties in the task of intercomparison. Simulations have uncertain forcings (Schmidt et al., 2011, 2012), and models contain parameterized or uncertain representation of the physics, chemistry, biology and interactions within the climate system (Flato et al., 2013). Furthermore, computational constraints impose a limited spatial resolution or a deliberate omission of some known processes in order to perform simulations at global scale that cover several centuries (e.g., Goosse et al., 2005; Schurer et al., 2013, Phipps et al., 2014)

The uncertainty in paleoclimatic reconstructions is not always well understood either and estimating its magnitude is challenging. For regional- to large-scale temperature reconstructions, uncertainty can be caused by random or systematic error in the proxy measurement, inadequate understanding of the proxy system response to environmental variation, differences in fields derived from instrumental records selected to calibrate the records, changes in the spatiotemporal and data type availability across the observational network, and reconstruction methods (e.g., Jones et al., 2009; Smerdon et al., 2010; Smerdon, 2012; Emile-Geay et al., 2013; Evans et al., 2013; Wang et al., 2014; Comboul et al., 2015).

The non-climatic noise in reconstructions has a significant influence on model data comparison. This may first have an impact on the variance of the reconstructed climatic signal itself, although this is dependent on the actual choice of calibration method (e.g., Hegerl et al., 2007; Christiansen et al., 2009; Mann et al., 2009; Smerdon et al., 2010; Smerdon, 2012). Furthermore, the non-climatic noise can mask real relationships between climate variations in different regions, or obscure the responses to forcing, which are clearer in models because of the absence of this noise.

Acknowledging the considerable uncertainty in paleoclimatic reconstructions, the earliest comparisons of past millennium simulations and reconstructions focused on hemispheric- and global-scale changes, using a single, often simple, climate model driven by globally uniform external radiative forcing estimates (e.g., Crowley, 2000; Bertrand et al., 2002). Later, simulations with more comprehensive models (e.g., Gonzalez-Rouco et al., 2006; Amman et al., 2007; Tett et al., 2007) refined the conclusions reached previously and enabled regional and continental-scale analyses. They underscored the potential role of the spatial distribution of some forcings, such as land use and of the dynamic response of the atmospheric circulation (e.g., Luterbacher et al., 2004; Raible et al., 2006; Goosse et al., 2006; Hegerl et al., 2011). Changes in the latter may be driven by the forcings (e.g., Shindell et al., 2001; Mann et al., 2009) or be a signature of internal variability of the climate system (e.g., Wunsch, 1999; Raible et al., 2005).

State-of-the-art climate models reasonably simulate properties of internal variability, such as teleconnection patterns or the probability of a particular event (e.g., Flato et al., 2013). However, they are not expected to reproduce the part of the observed time trajectory that is not directly constrained by external forcing because of the non-linear, chaotic nature of the system (Lorenz, 1963). This makes model-data comparison a complex issue when using a single simulation, because differences between model results and reconstructions may be due to a model or reconstruction bias, but may also simply reflect a different sample of internal variability (defined here as the fraction of climatic variability that is not due to changes in external forcings).

Indeed, comprehensive climate models have their own internal climate variability and, if a model represents the real world in a satisfactory way, the observed trajectory would just be one among all potential model realizations. The issue may be addressed by analysing an ensemble of simulations, which provides information on the range that can be simulated by one single model (e.g., Goosse et al., 2005; Yoshimori et al., 2005; Jungclaus et al., 2010; Moberg et al., 2015) or a set of models (e.g., Jansen et al., 2007; Lehner et al., 2012; Fernandez-Donado et al., 2013, Bothe et al., 2013b). The reconstruction has then to be compatible with this range, at least when considering all the uncertainties, to claim consistency between simulations and reconstructions, whereby such a compatibility can be defined in various ways, as discussed below.

Fernandez-Donado et al. (2013) reviewed results from 26 climate simulations with eight atmosphere-ocean general circulation models (AOGCMs), reflecting the state of modelling before the CMIP5/PMIP3 (Coupled Model Intercomparison Project Phase 5/Paleoclimate Model Intercomparison Phase 3). These pre-CMIP5/PMIP3 simulations were driven by a relatively wide range of choices for boundary conditions and forcing agents. For the Northern Hemisphere surface temperature variations, Fernandez-Donado et al. (2013) found an overall agreement within the temporal evolution, but still noted discrepancies between simulations and hemispheric and global temperature reconstructions. For example, the period between around 850 and 1250 CE is warmer in the reconstructions than in the simulations (see also Jungclaus et al., 2010; Goosse et al., 2012b; Shi et al., 2013).

Additionally, a comparison of the simulated changes in the temperature fields from this warm period and the colder period around 1450-1850 showed little resemblance to the field-reconstruction by Mann et al. (2009), but the spatial reconstructions themselves have significant uncertainties (e.g., Wang et al., 2015). These two relatively warm and cold periods are often referred to as the Medieval Climate Anomaly (MCA), and the Little Ice Age (LIA), respectively, although their exact timing has been debated and the adequacy of their names has been questioned (e.g., Jones and Mann, 2004; PAGES 2k Consortium, 2013).

The assessment of information from paleoclimate archives (Masson-Delmotte et al., 2013) in the IPCC fifth assessment report partly followed the approach applied by Fernandez-Donado et al. (2013). Masson-Delmotte et al. (2013) included a preliminary analysis of the more recent CMIP5/PMIP3 "past1000" simulations, which were coordinated more closely than previous experiments, particularly in regard to the choices of forcings (Schmidt et al., 2011, 2012). They came to similar conclusions as Fernandez-Donado et al. (2013): the reconstructed MCA warming is greater than simulated, but not inconsistent within the large uncertainties.

Agreement between paleoclimate reconstructions and simulations has also been assessed by compositing the response to individual forcing events (e.g., Hegerl et al., 2003; 2011; Luterbacher et al., 2004; Stenchikov et al., 2006; Masson-Delmotte et al., 2013). The reconstructed and simulated response to volcanic forcing agrees in magnitude on multi-decadal time scales. Detailed comparisons of observations around the 1815 Tambora eruption indicate that the simulated cooling is larger than in instrumental observations or in reconstructions (Brohan et al., 2012), but a significant part of the discrepancy might be due to forcing uncertainties.

For the solar forcing, direct comparisons between simulations and reconstructions are inconclusive regarding whether simulations that use either moderate or weak variations of total solar irradiance provide generally better agreement with reconstructions (Masson-Delmotte et al.; 2013; Fernandez-Donado et al., 2013). This has been confirmed at hemispheric and regional-scales by Hind and Moberg (2013) and Moberg et al. (2015), using appropriately designed statistical tests of temporal correlation and quadratic distance between reconstructions and simulations (Sundberg et al., 2012).

The cause of past climate change in the Northern Hemisphere, specifically the contribution by individual forcings to a climatic event, can be estimated using detection and attribution techniques. These techniques allow for the possibility that the reconstructions contain forced signals of larger or smaller magnitude than simulated (e.g., due to forcing uncertainty, uncertainty in a models transient response, or uncertainty in calibration of reconstructions). The results show that the response to volcanic eruptions can be clearly detected in reconstructions, consistent with epoch analysis results, and also confirm that the signal is generally larger in magnitude in the simulations (Hegerl et al., 2003; 2007; Schurer et al., 2013), although the discrepancy may be within the range of volcanic forcing uncertainty. The response to solar forcing cannot be reliably separated from internal variability, but very high solar forcing such as that reconstructed by Shapiro et al. (2011) needs to be significantly scaled down to match reconstructions even given large reconstruction uncertainties (Schurer et al., 2014). Within the LIA, detection and attribution methods show that volcanic forcing is critical for explaining the anomalous cold conditions (Hegerl et al., 2007; Miller et al., 2012; Lehner et al., 2013; McGregor et al., 2015) and that there is also weak evidence for a contribution from a small but long-lived decrease in $CO_2$ concentration (e.g., MacFarling-Meure et al., 2006; Schurer et al., 2014).

The studies mentioned above mainly focused on the Northern Hemisphere, because a larger number of paleoclimatic observations and reconstructions are available there. However, several recent studies assessed differences in inter-hemispheric connections (Goosse et al., 2004; Neukom et al., 2014), Southern Hemisphere climate variability (Phipps et al., 2014), regional temperature variability (Luterbacher et al., 2004; Hegerl et al., 2011; Goosse et al., 2012a; Gergis et al., 2015; Shi et al., 2015) and Southern Hemisphere circulation features (Wilmes et al., 2012; Abram et al., 2014; Tierney et al., 2015).

In particular, the recent consolidation of Southern Hemisphere paleoclimate data (Neukom and Gergis, 2012) led to the comparison of a hemispheric temperature comparison with a suite of 24 climate model simulations spanning the past millennium (Neukom et al., 2014). This study reported considerable differences in the 1000-year temperature reconstruction ensembles from the Northern and Southern Hemispheres. An extended cold period (1590s–1670s CE) was observed in both hemispheres, while the current (post-1974) warm phase is found to be the only period of the past millennium where both hemispheres experienced simultaneous warm anomalies (Neukom et al., 2014). Their analyses also suggested that the simulations underestimate the influence of internal variability in the ocean-dominated Southern Hemisphere (Neukom et al., 2014).

While several studies have provided valuable advances in our understanding of hemispheric-scale climate dynamics, this brief overview indicates that observed and simulated paleoclimate variations at regional and continental scales have not been thoroughly compared up to now. This was the goal of a workshop joining the PAGES 2k and PMIP3 communities in Madrid (Spain) in November 2013, using a recent set of continental-scale temperature reconstructions (PAGES 2k Consortium, 2013) and a collection of state-of-the-art model simulations driven by realistic external forcings (Schmidt et al., 2011, 2012). On the basis of the discussions held during this workshop, the aim of this study is to systematically estimate the consistency between the simulated and reconstructed temperature variations at the continental scale and evaluate the origin of observed and simulated variations. This study is motivated by the following key science questions:

1/ Are the statistical properties of surface temperature data for each individual continent-scale region consistent between simulations and reconstructions?

2/ Are the cross regional relations of temperature variations similar in reconstructions and models?

3/ Can the signal of the response to external forcing be detected on continental scale and, if so, how large are these signals?

Section 2 first presents a brief overview of the PAGES 2k reconstructions and simulations analysed here. In addition to a selection of PMIP3 simulations, some numerical experiments that did not follow the PMIP3 protocol were also analysed, mainly to include model runs with larger solar forcing amplitude. We use several statistical methods to achieve robust results in answering the key science questions above. They are listed at the end of section 2. Each methodology is briefly described when it is applied while some specific implementation information is provided in supplement sect. S2. In section 3, each continental-scale region is studied separately to determine whether the reconstructed and simulated time series have similar characteristics, in terms of the magnitude and timing of the observed changes as well as the spectral distribution of the variance. Section 4 investigates whether the inter-regional patterns of temperature variability are similar in the reconstructions and simulations. The role of the external forcings in producing the observed variations is presented in section 5. Section 6 provides a discussion of our results, their limitations and how our conclusions compare to previous studies. Finally, section 7 summarizes the main findings and provides perspectives for future developments. Several additional analyses are provided as supplementary material for completeness and further reference.

# 2. Data and methods

## 2.1 PAGES 2k reconstructions

The PAGES 2k Consortium (2013) generated temperature reconstructions for seven continental-scale regions (Fig. 1). The proxy climate records found to be best suited for reconstructing annual or warm-season temperature variability within each continental-scale region were identified. Expert criteria for the adequacy of proxies were *a priori* specified (PAGES 2k Consortium, 2013). The resulting PAGES 2k dataset includes 511 time series from different archives including tree rings, pollen, corals, lake and marine sediment, glacier ice, speleothems, and historical documents. These data record changes in biological or physical processes and are used to reconstruct temperature variations (all data are archived at: https://www.ncdc.noaa.gov/cdo/f?p=519:2:0::::P1_study_id:12621).

The PAGES 2k reconstructions have annual resolution in all regions except North America, which has one 780-year-long tree-ring-based reconstruction (back to 1200 CE with 10-year resolution) and one 1400-year-long pollen-based reconstruction (back to 480 CE with 30-year resolution). These latter two reconstructions therefore are smoothed differently and they are either excluded from the analysis or treated in slightly different ways in some comparisons. The reconstruction for the Arctic region used in this study is based on a revised version (v 1.1) of the PAGES 2k dataset (McKay and Kaufman, 2014).

Each regional group tailored its own procedures to their local proxy records and regional calibration targets (PAGES 2k Consortium, 2013). Thus, each continental-scale temperature reconstruction was derived using different statistical methods. In short, most groups used either a scaling approach to adjust the mean and variance of a predictor composite to an instrumental target, or a regression-based technique to extract a common signal from the predictors using principal components or distance weighting. Thus, some of the observed region-to-region differences between simulations and reconstructions might be due to the differences in reconstruction methods. Nevertheless, alternative reconstructions for all regions based on exactly the same statistical procedures were also produced and were found to be similar to the PAGES 2k temperature reconstructions provided by each group (PAGES 2k Consortium, 2013). Each regional group also used individually selected approaches to assess the uncertainty of their temperature reconstructions, designed to quantify different aspects of the uncertainty. For example, some regions primarily quantified uncertainties associated with the set of records used in the reconstruction and their agreement through time, which can reflect within-region variability as well as uncertainty (Arctic, North American tree rings). Other regions focused on uncertainties associated with how closely the proxy resembles temperatures (Asia, Antarctica, Europe, North American pollen), and some regions incorporated both of these types of uncertainties (Australasia, South America). Uncertainty estimates in all of the regions except for Antarctica vary through time depending on the set of records available for any given interval and their agreement. All uncertainty estimates that assess how well the proxy data reproduce observed temperatures are based on the assumption that the modern proxy-temperature relation is stationary into the past, and that the agreement between proxy data and temperature on short timescales can be used to infer uncertainty at lower frequencies.

### 2.2 Climate model simulations

The climate model simulations used in this study are listed in Table 1, summarizing model specifications such as resolution, forcing applied to the transient simulations, and length of pre-industrial control simulations (piControl). These simulations include contributions to the third Paleoclimate and the fifth Coupled Modelling Intercomparison Projects (PMIP3, Braconnot et al., 2012; CMIP5, Taylor et al., 2012) from six models (CCSM4, CSIRO-Mk3L-1-2, GISS-E2-R, HadCM3, IPSL-CM5A-LR, MPI-ESM-P), as well as a more recent simulation with CESM1, and the COSMOS pre-PMIP3 ensemble with ECHAM5/MPIOM (see also Table S1).

The experiments were selected among available pre-PMIP3 and PMIP3 simulations on the basis of specific criteria: The conditions were (i) they run continuously from 850-2000 CE; (ii) they include at least solar, volcanic aerosol, and greenhouse gas forcing (S, V, G in Table 1); (iii) they use a plausible solar forcing reconstruction with an amplitude within the range that is consistent with recent understanding (iv) they do not display a large unphysical drift over the simulated period.

PMIP3 simulations all comply with criteria (ii) and (iii) as they follow the recommendation of Schmidt et al. (2011) by using an increase in total solar irradiance (TSI) from the late Maunder Minimum period to the present day of ~0.10%. Nevertheless, some PMIP3 simulations were excluded from the analysis, as the simulations presented clear incompatibilities with the rest of the ensemble. For instance, the MIROC simulation displays a trend in the global annual mean temperature over the whole millennium that is not compatible with the present understanding of the past millennium climate. It has been considered here as a likely model artefact that could also affect regional and seasonal temperatures in unknown ways. Contrary to the GISS model, this drift is not clearly understood and no control run is available to statistically correct it. The simulation with bcc-csm-1 was discarded because of potentially unphysical large anomalies in some regions. FGOALS-gl was not use due to the unavailability of a continuous run from 850 to 2000, as the so-called 'past1000' simulation covers only the years 850-1850 under the PMIP protocol.

Most non-PMIP3 simulations did not comply with at least one the criteria above. Nevertheless, experiments performed with two models (ECHAM5/MPIOM and CESM1) follow all of them. They include simulations with a stronger solar forcing than in the PMIP3 ensemble. A three-member ensemble with ECHAM5/MPIOM uses a TSI reconstruction with an increase of ~0.24% (COSMOS E2), while CESM1 uses a TSI reconstruction with an increase of ~0.20%. No simulation used in this study incorporates the much larger increase of ~0.44%, suggested by Shapiro et al. (2011), which results in simulations that are inconsistent with reconstructed large-scale temperatures (Feulner, 2011; Schurer et al., 2014). The COSMOS simulations deviate from the PMIP3 protocol because they included an interactive carbon cycle with $CO_2$ concentration as prognostic variable. While simulated and reconstructed $CO_2$ evolution diverge during some periods, the differences have only a marginal effect on simulated temperatures (Jungclaus et al., 2010).

Consequently, the group of simulations analysed here is not strictly based on the PMIP3 ensemble. Nevertheless, as we use a majority of PMIP3 simulations and additional simulations that follow an experimental design similar to PMIP3, we will keep the reference to PMIP3 for simplicity.

The variable extracted from the simulation outputs is the monthly mean surface air temperature (labelled ‘tas’ in the Climate Model Output Rewriter framework of CMIP5). These temperature fields were then used to create area-averaged time series that matched the domain and seasonal window of each of the PAGES 2k regional reconstructions (see Supplement Sect. S1).

### 2.3 Statistical methods

Several climate model–paleoclimate data comparison and analysis methods are used in this study to verify the robustness of the results generated by each method and to provide a comprehensive guide for future work. Model-data comparisons need to account for uncertainties in climate reconstructions, in forcing reconstructions and in the response to forcings in model simulations. These approaches also must recognize that the real climate, and hence the reconstructions, and individual climate model simulations include their own individual realizations of internally generated variability. Therefore, perfect agreement between model simulations and data can never be expected when directly comparing time series.

The first group of methods is focused on the first question raised in the introduction. The goal is to assess if temperature reconstructions have similar statistical properties compared to simulations. This is initially done by simple analysis of the time series, such as estimates of the variance (section 3.1). The spectral properties are then analysed (section 3.2) before the probabilistic and climatological consistency (section 3.3) and the skill of the various simulations (section 3.4). The second question dealing with the cross regional variations of temperatures is addressed by discussing the correlation between regions (sections 4.1 and 4.3) and through a principal component analysis (section 4.2). Finally, the third question about the role of the forcing is studied by means of a superposed epoch analysis (section 5.1), by applying a statistical framework involving correlation and distance metrics (section 5.2) and detection and attribution techniques (section 5.3). For more details on those methods, see supplement sect. S2.

In the majority of the analyses presented in this manuscript, anomalies compared to the mean over the whole period covered are used and the time series are smoothed or temporally averaged, using either a 23-point Hamming filter or non-overlapping 15-year averages, depending on the requirements of the various techniques (both methods give a similar degree of low-pass filtering). This is motivated by the relatively weaker skill of some reconstructions to replicate observed records on interannual time scales (Cook et al., 2004; Esper et al., 2005; D’Arrigo et al., 2006) and by the fact that the main focus here on decadal to centennial timescales. The full period analysed is 850-2005 CE, although different periods are chosen for some analyses because of data availability, the choice of the temporal filtering, other technical restrictions, or to analyze sub-periods.

## 3. Regional analysis

To begin, the agreement between simulations and reconstructions for individual regions is described qualitatively, using a simple visual comparison of the time series, and then quantitatively by calculating spectra, consistency and skill metrics. The correlations between the time series are presented in the supplementary material (Figure S1 and supplementary text S3). Overall, the analyses in this section illustrate the potential of identifying common signals in both data sets. The different diagnostics are presented here separately whereas the conclusions derived from the results of the different analyses are compared and discussed in more detail in section 6.

### 3.1 Observed and simulated time series

Figure 1 shows the regional time series in the forced simulations with each regional temperature reconstruction. To the right of each time series graph, the magnitude of variability of unforced simulated temperatures is illustrated by calculating the standard deviation of pre-industrial control simulations in each model. The unforced variability is generally similar in all models in all the regions, with weaker amplitudes in Australasia and Asia. Note that some regions cover only land areas while others have an oceanic fraction (see Supplement Sect. S1), with a potential impact on the magnitude of the estimated variability.

Most reconstructions show a tendency of a gradual cooling over the millennium, followed by recent warming. Notable common features among regions on decadal timescales are the pronounced negative anomalies related to large tropical volcanic eruptions in the simulations. This is most obvious for the eruptions in the 1250s, 1450s and 1810s. Among the temperature time series, a larger response to volcanic eruptions is noticeable in the CESM, MPI and CCSM4 simulations. The regional temperature reconstructions rarely capture the first two of these anomalies or only register them at smaller amplitudes. Only the early 19$^{th}$ century eruptions are clearly reflected in many regions, and are most pronounced in the Northern Hemisphere reconstructions. The reconstruction for Europe also shows a negative anomaly coinciding with the effect of the 1450s eruption, with an amplitude comparable to that seen in some of the simulations.

Figure 1 suggests that the temperature reconstructions show slightly more centennial to multi-centennial variability than the models over the full period with stronger long-term trends, while several model results indicate a stronger recent warming compared to some of the reconstructions. The reconstruction uncertainty bands provided with the original PAGES 2k reconstructions encompass the simulated series with few exceptions, in particular the Arctic and North America during the 1250s. The published uncertainty estimates have been calculated using different methods for the various continental-scale regions, as detailed in the supplementary material of PAGES 2k Consortium (2013). Furthermore, those uncertainties are only valid at the original temporal resolution, which is annual in all cases except for North America. It is expected that the reconstruction uncertainty decreases at lower resolution, or after smoothing as in our case. This is consistent with the lower uncertainty ranges for the low-resolution pollen-based reconstruction.

However, estimating the reduction of the uncertainty due to smoothing is not straightforward (e.g., Moberg and Brattström, 2011; Franke et al., 2013) as the resulting uncertainty magnitude is also dependent on autocorrelation of the non-climatic noise in proxy data. The extreme hypothesis, considering that the error is constant in time and that the errors are uncorrelated would lead to a decrease proportionally to 1 over the square root of the number of samples included in the average. For a smoothing similar to 15-year averaging, as performed herein, the approximation that likely leads to an underestimation of the uncertainties would correspond to a decrease by a factor of about 4 compared to the original error estimate. This suggests very small errors for most reconstructions. In this case, the major discrepancies between the reconstructions and model results would occur at the same time as mentioned above; however, periods when the models are out of the range of the reconstruction uncertainty bands would be more common at the decadal scale.

For North America, the long term multi-centennial trend appears to be similar between the pollen based reconstruction and simulations, except for the last ~200 years, when some simulations show much stronger warming than is present in the reconstruction. This warming feature is somewhat stronger in the tree-ring based reconstruction than in the pollen-based reconstruction, but is nevertheless weaker than in some simulations. The COSMOS simulations appear to be collectively colder than this reconstruction in the late 20$^{th}$ century. Although the European temperature reconstruction and simulated series disagree substantially in some parts of the 12$^{th}$ century and for the last ~200 years, there are otherwise strong similarities, particularly during periods of large volcanic eruptions. Simulated and reconstructed Arctic series show large decadal to centennial variability, but the timing of these variations do not agree well. Therefore, simulations are often outside the reconstruction's uncertainty range. Consistently, there is a large multi-model ensemble-spread but also single-model ensemble spread as illustrated by the COSMOS simulations. CESM, CCSM4, and IPSL show a strong recent warming and strong volcanic cooling.

Simulated and reconstructed temperatures show only weak long-term trends in Asia, but decadal variability appears to be larger in the reconstruction. Simulations generally differ from the reconstruction in the last 200 years and show either much weaker or much stronger trends. In Australasia, the weak forced variability common to all simulations may be due to the large spatial extent of the domain, which includes large oceanic areas that may dampen the forced high-frequency variability. For the recent warming, the trends in CESM, CCSM4, IPSL and the COSMOS simulations are considerably stronger than the Australasian temperature reconstruction. The temperature reconstruction for South America is often near the upper or lower limit of the simulation ensemble range and displays more centennial-scale variability than the simulations. In Antarctica, the reconstruction has a clear long-term negative trend and only a modest warming in the 20$^{th}$ century while the simulations show nearly no long-term cooling but agree on the warming onset in the beginning of the 20$^{th}$ century.

### 3.2 Spectral analysis

Next, we consider the agreement between simulated and reconstructed temperature data in terms of their spectral densities, which show how temperature variances are distributed over frequency (Fig. 2, see also Fig. S2). Spectra were computed using the multi-taper method (Thomson, 1982; Percival and Walden, 1993), with its so-called time-bandwidth product being set to four. Consequently, each calculated spectrum is an average of seven statistically independent spectrum estimates. Spectra for the reconstructions are illustrated with their 95% confidence intervals, while model spectra are plotted with single lines. The analysis is made at the original time resolution using all existing data points in the time window 850-2005.

The degree of agreement between model and reconstruction spectra differ substantially between regions, with the Arctic showing the best agreement at all frequencies and South America showing the worst. In the latter, most model spectra lie in the reconstruction confidence interval only in a narrow frequency band corresponding to about 100- to 150-year periods. The agreement is generally good for the Arctic, Europe and Asia and at multi-decadal timescales (20-50 years) for many regions. Nevertheless, many models have systematically less variance in the 50- to 100-year band and most models have more variance than the reconstructions at higher frequencies.

Pronounced differences of high-frequency variance is seen for all Southern Hemisphere regions. In particular, the pre-PMIP3 COSMOS simulations show significantly too much variance at timescales of 3 to 5 years for Australasia and to a lesser degree for South America and Antarctica. This property has previously been related in regions with strong influence from tropical Pacific variability to this model's ENSO variability (Jungclaus et al., 2006; Fernández-Donado et al., 2013). Most model spectra for North America lie within the confidence interval of the tree-ring based reconstruction spectrum, although several models have somewhat less variance than this reconstruction at periods longer than 50 years. The North America pollen-based reconstruction behaves as a roughly 150-year low-pass filtered series and has significantly less variance than the corresponding tree-ring-based record at all frequencies for which both spectra are defined.

### 3.3 Consistency estimate

The probabilistic and climatological consistency of PMIP3 simulations and PAGES 2k reconstructions was assessed following the framework of Annan and Hargreaves (2010; and references therein; Hargreaves et al., 2011, 2013) and Marzban et al. (2011), respectively. The current application is based on Bothe et al. (2013a,b). The underlying null hypothesis follows the paradigm of a statistically indistinguishable ensemble (Annan and Hargreaves, 2010; Rougier et al., 2013), i.e. the validation target, represented here by the temperature reconstructions, and the model simulations are samples from a common distribution and are therefore exchangeable.

Climatological consistency refers to the similarity of the climatological probability distributions of reconstructions and of simulations over a selected period, either the whole millennium or sliding sub-periods. We analyse climatological consistency by comparing individual simulated series with the target (i.e., the reconstructions) to identify deviations in climatological variance and possible biases between them. To achieve this goal, Marzban et al. (2011) proposed the use of residual quantile-quantile (r-q-q) plots that should be approximately flat for consistent series (supplement sect S2.1).

Probabilistic consistency refers to the position of the reconstruction in the range spanned by the ensemble of simulations. Histograms of the ranks should be flat under exchangeability (supplement sect S2.1), i.e. estimated frequencies of the verification target and the ensemble agree if the simulation ensemble is probabilistically consistent with the temperature reconstructions (Murphy, 1973).

As there are large uncertainties in paleoclimate reconstructions, it is necessary to take into account these uncertainties in the evaluation of the consistency of the ensemble of climate model simulations (Anderson, 1996). This is achieved by inflating the model simulations results by adding noise with amplitudes that are proportional to published uncertainty estimates from the original temperature reconstructions.

We assess probabilistic and climatological consistency based on non-overlapping 15-year averages centred on the full period considered, except for the Northern American temperature reconstruction where non-overlapping 30-year averages are used for the pollen-based reconstruction, and 10-year averages for the tree-ring-based reconstruction. The results are presented on Figure 3, Figure S3 and Figure S4 for all regions.

The regions selected for Figure 3 are chosen to provide a contrasting example. Two estimates of the uncertainties are used. First, the uncertainties provided with the original reconstruction are applied, which is an overestimation for the smoothed time series. Second, at the other extreme, the uncertainties are assumed to be equal to zero and are thus known to be underestimated. A third estimate of the uncertainty is provided in the supplementary figures, using an uncertainty measure equal to the one provided in the original publication divided by a factor $\sqrt{15}$ to account for the smoothing (see section 3.1). This leads to results that are generally very similar to the case where uncertainty is assumed to be zero.

The simulations in most cases lack climatological consistency with the reconstructions (Fig. 3 and Fig. S3). The simulated quantiles can deviate strongly from the reconstructed quantiles. Specifically, the simulated distributions are generally over-dispersive when using the original estimates of uncertainties. The differences are much smaller when uncertainties in reconstructions are neglected, although extremes often remain overestimated. The Arctic and the North American tree-ring based reconstruction are exceptions as some simulations are climatologically consistent with the reconstruction and display only small differences between simulated and reconstructed quantiles for all estimates of the uncertainty. Consistency is reduced for those simulations that show larger variability (recall Figure 1) as is the case of the CCSM4 and CESM models.

In agreement with the climatological assessment, the simulated results generally lack probabilistic consistency with the reconstructions when the original uncertainty is considered (Fig. 3 and Fig. S4). The target data are too often in the central ranks, indicating that the probabilistic distribution of the ensemble is too wide and shows significantly over-dispersive spread deviations. The only exception is the North American region using the tree-ring based reconstruction. The most prominent differences are found in the Antarctic region where the simulation ensemble spread deviates considerably from reconstructed temperatures (Fig. S4), but strong ensemble spread deviations relative to the pollen reconstruction for North America are also evident.

This assessment of the probabilistic consistency strongly depends on the estimate of the uncertainty of the reconstruction. If we do not add noise to the model time series to reflect error in reconstructions before the ranking and thereby neglect reconstruction uncertainty, or if we assume a strong reduction of the error in reconstruction at the decadal time scale because of the smoothing, the ensemble appears to be consistent with a number of regions or even under-dispersive for others. However, ignoring the uncertainty in such a manner may lead to an overconfident assessment of consistency between simulation ensemble and reconstruction. Nevertheless, because the uncertainties are not well known, over-dispersion does not necessarily weaken the reliability of the ensemble relative to the target, but instead may highlight insufficiently constrained uncertainties in the reconstruction.

### 3.4 Skill estimate

The skill of the simulations is assessed using a metric introduced by Hargreaves et al. (2013). The idea of skill stems from weather forecasting and refers to the ability of a simulation to represent a target better than some simple reference values. For instance, in weather forecasting, a standard reference is to assume no change compared to initial conditions (i.e., persistence). A forecast has a positive skill if it is closer to the observed changes than this simple reference. The skill *S*, as in Hargreaves et al. (2013), is then:

$$S = 1 - \sqrt{\frac{\sum (F_i - O_i)^2 - \sum e_i^2}{\sum (R_i - O_i)^2 - \sum e_i^2}} \quad (1)$$

where $F_i$ is the simulation result at each data point, $O_i$ is the reconstruction data, $R_i$ is the reference (for instance a constant climate here) and $e_i$ is uncertainty of the target. The square-root expression becomes undefined when either the actual simulation or the reference is better than the upper possible agreement level indicated by the errors. Uncertainty estimates are derived from the originally reported uncertainties in regional temperature reconstructions given by PAGES 2k Consortium (2013). If reconstructed error estimates are realistic, we do not expect the simulations to fit the target better than these uncertainty estimates. As for the consistency analyses, the skill analysis is calculated using temperature anomalies from the long-term averages within each analysis period.

Figure 4 presents the skill for the Arctic and Antarctica, as an example, with the other PAGES 2k regions displayed in Figure S5. In this estimate, we use a no-change reference forecast (i.e., the reference is the climatology) as there is no clear *a priori* evidence that the climate at one particular time during the past millennium is warmer or colder than the mean. Positive values suggest that the simulations is in better agreement with (i.e., closer than) the regional reconstructions than this reference. Results are presented for dates when no data are missing in four periods: 850 to 1350, 1350 to 1850, 850 to 1850, and the full period 850 to 2000. As in section 3.3, we compute the skill in Fig. 4 using the uncertainties provided with the original reconstruction, as well as a case that assumes the uncertainties are negligible (i.e., assuming $\sum e_i^2 = 0$ in Eq. 1 of section 3.4). Additionally, the skill is computed assuming a reduction by a factor $\sqrt{15}$ in the supplementary figures.

The most notable result is that the skill measure is generally undefined when using the uncertainties provided with the original reconstruction: either the reference or the simulated data are closer to the reconstruction than uncertainty allows, leading to the square root of a negative number in Eq. 1. This confirms that uncertainties in the reconstructions are potentially an overestimation for smoothed time series. When ignoring uncertainties, the 15 year non-overlapping means of the simulations rarely display skill. Simulation skill appears to be most likely for the European and Arctic regions, while positive skill is nearly absent for the Southern Hemisphere regions and North America in all the models.

## 4. Links between the different regions

The structure of the spatial variability, i.e. the spatial covariance of temperature changes, contains contributions from forced signals and from teleconnections in the internal climate variability. The PAGES 2k temperature reconstructions help to investigate the consistency between simulations and reconstructions with respect to this covariance structure. In the following sections, this is evaluated using spatial correlations, Principal Components (PCs) and Empirical Orthogonal Functions (EOFs), and correlations over sliding temporal windows.

### 4.1 Spatial correlation

The spatial correlation matrix of simulated temperature for the PAGES 2k regions is compared to the correlation matrix of the PAGES 2k reconstructions (Fig. 5 and Fig. S6). Correlations are calculated for detrended continental mean time series filtered with a 23 year Hamming window and based on the continents for which these are available, which excludes North America. We use the longest common period for forced simulations and reconstructions, which for the filtered data is 1012 CE – 1978 CE (1000 CE – 1990 CE for annual data). To disentangle the contributions from forcings and from internal variability we analysed forced simulations for the entire analysis period, forced simulations for the pre-industrial period (before 1850 CE), and unforced control simulations.

MPI-ESM-P is used to illustrate our main findings in Figure 5 (see Fig. S6 for the other models). Correlations in the forced MPI-ESM-P simulation for the whole period are higher than 0.6 between nearly all regions. In contrast, the correlations for the PAGES 2k temperature reconstructions are rather low, which indicates a substantial inconsistency between the correlation structure in the models and in the PAGES 2k temperature reconstructions. The potential causes of this discrepancy will be discussed in section 6 but we must recall here that, in contrasts to other analyses presented above, the evaluation of the spatial correlation does not take into account any uncertainty in the reconstruction. Any non-climatic noise related to the characteristics of the proxy records selected or differences in the reconstruction method between regions would decrease the correlation, contributing to have lower values than for the model results.

The correlations in the simulation are lower if only the pre-industrial period is considered, and close to zero in the control simulations. The simulated high correlations for the last century are likely to be a consequence of the rather homogeneous and strong anthropogenic warming in the simulations. The high correlations for the pre-industrial forced runs show that the response to volcanic forcing, solar forcing land use and/or orbital forcing also substantially contributes to the correlations at the time scales considered. Low values obtained for the control simulations indicate that teleconnections between continents are weak for simulated internal variability.

Although these general characteristics are present in many of the models evaluated here, there are some differences among them. In particular, some of the models that show higher correlations during pre-industrial times (e.g., CESM) also display a large response to volcanic forcing compared to the other members of the ensemble (Lehner et al., 2015). Additionally, the specific characteristics of some regions may differ substantially. For instance, the correlation between Antarctic temperatures and other regions is very low in MPI-ESM-P or IPSL-CM5A-LR for pre-industrial conditions while it is much larger in CCSM4 and CESM. This can be attributed to a different ratio of forced versus unforced variability, and in particular to discrepancies in the magnitude of the response to external forcing in the selected models.

### 4.2 Principal Component Analysis

Figure 6a shows the loadings of the first EOF on each region for the PMIP3 forced simulations and the PAGES 2k reconstructions (with corresponding results for the GISS and COSMOS ensembles presented in the supplementary text S4 and Fig. S7). Most models show similarities in the loadings, which indicates that the different regions co-vary similarly in the different models. All loadings are positive and thus the first principal component (PC) is a just weighted mean of all continental temperature series.

Consequently, the time series of the first PC of the PMIP3 simulations and PAGES 2k temperature reconstructions (Fig. 6b) reflect the main features of the individual original series (particularly for Northern Hemisphere regions); namely a temperature decline after around 1200 CE, which lasts until the early 1800s, followed by the sustained warming within the 19$^{th}$ and 20$^{th}$ century. Additionally, the influence of volcanic eruptions on reconstructed temperatures is visible during some periods, especially during the mid-13$^{th}$ century (although not in the reconstructions), the mid-15$^{th}$ century and the beginning of the 19$^{th}$ century.

In most models, the first EOF explains about 80-90% of the total variance, whereas the leading EOF in the PAGES 2k temperature reconstructions accounts for only 55% of the total variance. This shows that the covariance structure is less complex in the simulations. This is consistent with the larger correlations between regions found in section 4.1, which means that the leading mode of homogeneous warming or cooling dominates the covariance structure in model results. In a few simulations (HadCM3, COSMOS), however, the variance explained by the first EOF is about the same as in the reconstructions.

The largest values for the loadings are found for the Arctic region, due to the high temperature variability in the last 1,200 years in this region. This expression of the classical Arctic amplification is reflected in most models and in the reconstructions. The ocean-dominated regions of the Southern Hemispheric show less pronounced variability relative to the Northern Hemisphere, consistent with the results of Neukom et al. (2014).

If the analysis is performed over the pre-industrial period only (Fig. S8), similar conclusions are reached but the loadings are smaller, especially over the Arctic, and the amount of variance represented in the leading EOF generally decreases, indicating a larger heterogeneity in the pre-industrial period.

### 4.3 Inter-regional and -hemispheric coherence of past temperature variability

Next, the stationarity of the correlation structure between the different regions, in the models and the reconstructions, is assessed using a running correlation analysis, (Fig. 7, Fig. S9). For the simulations, the multi-model mean shows generally high inter-regional correlations, as the common contribution of the forcing is enhanced because of the averaging procedure. Periods with small variations in external forcing are, however, characterized by weaker coherence between the regions. This occurs during the 11$^{th}$ and 12$^{th}$ century and in shorter periods around 1500 and 1750. High coherence occurs in periods with strong variations in external forcing, highlighting in particular that volcanic eruptions can cause simultaneous temperature variations in most regions.

The inter-regional correlations in the individual model simulations vary considerably. The model range includes the correlations derived from the reconstructions for some regions, as for Europe vs. Arctic (Fig. 7a), but values for models are very often higher than for reconstructions (see also section 4.1). The difference is particularly large for the coherence between Australasia and South America (Fig. 7b), which is substantially larger in model simulations compared to reconstructions and instrumental observations (Morice et al., 2012) (Fig. 7b). This could indicate that some regions are less connected by modes of variability (such as ENSO) in reality than suggested by models, that the models have poor representation of modes of internal variability that influence the ocean-dominated Southern Hemisphere (see Neukom et al., 2014; see also Supplementary text S5 and Fig. S10), or that there is more non-temperature noise in the proxy data from those regions.

## 5. Role of Forcing

Some aspects of the response to external forcing have been briefly discussed in the previous sections. It is now formally addressed here by a superposed epoch analysis, by applying the $U_R$ and $U_T$ (correlation- and distance-based) model evaluation statistics and by detection and attribution techniques.

### 5.1 Superposed epoch analysis

The response of the PAGES 2k reconstructions and the various model simulations to external forcing from solar and volcanic activity is evaluated here using a superposed epoch analysis approach, following Masson-Delmotte et al. (2013). This analysis was conducted for two different time scales, interannual and multidecadal. For interannual time scales, this is done by generating composites of reconstructed and simulated temperature sequences corresponding to the timing of the 12 strongest volcanic events (see supplement sect S2.2). 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 from time series smoothed with a 40-year low pass filter using least-squares coefficients (Bloomfield, 1976). We also calculate composites corresponding to the timing of intervals of weaker solar forcing at decadal timescales. The intensity of the average model response to the selected forcing events is then compared to the corresponding response found in the reconstructions.

The regional temperature response for six PAGES 2k regions (North America is not analysed here; see Section 2.1) to the major volcanic events in the Crowley and Unterman (2012) reconstruction are shown in Figure 8. The temperature perturbation typically lasts longer than the forcing itself, with a recovery to pre-eruption temperatures after 3 to 10 years in the simulations and in the reconstructions.

The responses vary considerably in the simulations and in the reconstructions among regions. Nevertheless, the composite averages are always larger in model results with values of up to -1°C compared to about -0.25ºC in reconstructions. The largest responses in simulated and reconstructed temperatures are found in Europe and Asia. The Arctic and South America show smaller simulated temperature changes compared to Europe and Asia (around -0.5ºC) and the average responses in the reconstructions are even smaller but stay at levels of -0.1 to -0.2ºC during several years. For the Antarctic region, both the simulated and reconstructed temperature response is negligible. This is also the case for the reconstructed response in Australasia. Similar results were obtained using the Gao et al. (2008) forcing (see supplementary text S6 and Fig. S11).

At multidecadal timescales, the simulated and reconstructed temperature responses are in better agreement, in particular when using the Crowley and Unterman (2012) reconstruction (Fig. S12) rather than the Gao et al. (2008) estimations (Fig. S13), with temperature decreases on the order of a few tenths of degree in most regions. The one exception is South America where, in contrast to simulations, the reconstructions do not show any multidecadal changes associated with volcanic forcing.

The multidecadal impact of solar forcing in the reconstructions is strongest in Europe, the Arctic and Asia (Fig. S14), with mean changes ranging from 0.15 to 0.25ºC. Changes in model simulations are smaller, lying between 0.05 and 0.1ºC in all regions except for

Antarctica where no changes are perceptible. The reconstructed changes thus appear larger than the simulated ones in Europe and the Arctic. This interpretation of the results should be approached cautiously, however, as the solar variability is not independent from the volcanic forcing analysis. Volcanic eruptions tend to occur more often in periods of low solar forcing in the reconstructed forcing records, and solar forcing itself is characterised by significant uncertainties (e.g., Schmidt et al., 2011).

### 5.2 Framework for evaluation of climate model simulations: $U_R$ and $U_T$ statistics

A statistical framework for evaluating simulated temperature sequences against reconstructed past temperature variations was developed by Sundberg et al. (2012), Hind et al. (2012) and Moberg et al. (2015). It includes two essential metrics, which both serve as statistical tests of a null hypothesis. First, a correlation metric, $U_R$, is used to test whether a significant positive correlation exists between simulated and observed (or reconstructed) temperature variations, indicating that they share a common response to changes in external forcings.

Second, a weighted square-distance metric, $U_T$, is used to test whether temperature variations in a forced simulation are significantly closer to the observed temperature variations than an unforced control simulation. If this is the case, a negative $U_T$ is obtained, whereas a positive $U_T$ indicates that the simulated response to forcings is larger than those in the observations, provided a significant positive $U_R$ is found. Both metrics are approximately distributed as N(0,1) under the null hypothesis that forced simulations are equivalent to unforced control simulations. Thus, it is easy to see if a $U_R$ or $U_T$ value is significantly different from zero. For example, a one-sided test value numerically larger than 1.65 is significant at the 5% level.

Prior to the analysis, all records were recalibrated against their instrumental target temperature time series (see supplement sect S2.3) following the procedure of Sundberg et al. (2012) and Moberg et al. (2015) to ensure that each regional reconstruction, after calibration, approximately satisfies the assumption that the true temperature component, upon which the non-climatic noise component is added, is correctly scaled (see Sundberg et al., 2012 and Moberg et al., 2015).

Figure 9 shows the model evaluation statistics $U_R$ and $U_T$ (Sundberg et al., 2012), calculated for the 861-1850 pre-industrial period. In general, all forced simulations and the reconstructions share a common forcing signal and, overall the forced simulations match the reconstructions significantly better than the unforced control simulations. However, these overall positive results are essentially due to a good match between simulations and reconstructions in the Northern Hemisphere while the agreement is poorer in the Southern Hemisphere.

Because of the imprint of the forcing response, all forced simulations show significant ($p<0.01$) positive correlation statistics ($U_R$) when data from all seven regions are combined, although notable differences are seen between regions. In the Arctic, Europe and Asia, all simulations have significant positive $U_R$ values. Nearly all simulations for Australasia and most for Antarctica also have significant positive $U_R$.

In contrast, simulated and reconstructed pre-industrial temperature histories for South America show little common variation, as revealed by mostly insignificant $U_R$ (some are even negative) in that region. $U_R$ statistics for North America (tree-ring based reconstruction) are only slightly better, but note that this reconstruction only starts in 1201. Moreover, the original temporal resolution of 10 years in the North American reconstruction leads to some loss of information in this analysis, which was performed at a 15-year resolution.

Results for the distance statistic ($U_T$) show that nearly all forced simulations are significantly closer ($p<0.05$) to the observed temperature variations than their respective control simulations when all regions are combined, i.e. their $U_T$ statistics are negative and statistically significant. The Arctic shows the overall best performance in the sense of having the largest number of negative significant $U_T$ values. Most simulations also show negative $U_T$ for Europe, Asia and Antarctica, but many of them are insignificant. For Australasia and South America, nearly all $U_T$ values are insignificant and many are even positive.

Thus, overall, the comparison between simulation results and reconstructions performs notably better for the Northern Hemisphere than for the Southern Hemisphere. In particular, nearly all simulations have significant negative $U_T$ values for the combined Northern Hemisphere data ($p<0.05$) but no significant negative values are found for the Southern Hemisphere where most of the $U_T$ values are positive. This suggests that the simulated effect of forcings in the Northern Hemisphere agrees well in amplitude with the corresponding effect in the Northern Hemisphere reconstructions, whereas the simulated Southern Hemisphere effect of forcings often appears to be larger than in the Southern Hemisphere reconstructions.

Results for both $U_R$ and $U_T$ suggest that the most robust agreements are for the largest spatial scales and for ensemble mean results (Fig 9). The most significant $U_R$ and $U_T$ are for ensemble means and global comparisons, followed by ensemble means and NH comparisons. Splitting the analysis period into two halves (856-1350 and 1356-1850, Figs. S15-S16) shows that the more recent period has better $U_R$ statistics. There are, however, not many significant negative $U_T$ in this period, although North America in particular shows several significant values. Extending the analysis to the full 861-2000 period yields higher $U_R$ values for most regions (Fig. S17).

The exception is Antarctica, where lower $U_R$ values indicate a divergence of the simulations and reconstruction for this region during the industrial period. Notably, $U_T$ values for the full analysis period are mostly weaker than for the pre-industrial period. Consequently, the overall performance of the simulation results-reconstruction comparison degrades in terms of the distance measure when recent data are included. This is likely because the simulated signal itself often has a larger amplitude in the industrial period than many of the regional temperature reconstructions (see Fig. 1).

Ensemble means for COSMOS and GISS ensembles give more significant $U_R$ and $U_T$ than individual simulations from these ensembles, demonstrating once more the value of averaging for isolating the forced signal. The intra-ensemble spread of test statistics illustrates the degree of randomness in $U_R$ and $U_T$ statistics for individual simulations, highlighting the danger of judging one model as being better than another. In particular, it is difficult to judge whether the high (E2) or low (E1) solar forcing amplitude of the COSMOS simulations provides a better fit to the reconstructions, as their ranges of $U_T$ values for individual simulations mostly overlap. For the early period analysis (856-1350), however, the low solar COSMOS simulations provide a better fit than the high solar simulations, as seen by their respective $U_T$ values being of different sign and having entirely non-overlapping ranges when all seven regions or when the Northern Hemisphere regions are combined (Fig. S15).

This result is confirmed by a formal test where $U_T$ is calculated in a different way to directly compare the two COSMOS ensembles, using the method described in Moberg et al. (2015). This test reveals that, despite a significantly better fit of the low solar simulations in the earliest period, neither of the two solar forcing alternatives gives a significantly better fit to the reconstructed temperature history when the more recent data are included (Fig. S18).

### 5.3 Detection and attribution

Detection and attribution aims to identify the forced response in the regional temperature reconstructions by evaluating if observed changes could be entirely caused by variability created within the climate system (internal variability), if external forcing is necessary to explain them, and what magnitude of external forcing response is consistent with reconstructions (see Bindoff et al., 2013; Hegerl and Zwiers, 2011). Here, we focus on all forcings together and not on the response to each forcing individually, as simulations with individual forcings are needed to analyse the latter. Attribution is achieved by estimating the response to the external forcing in the reconstruction using a total least squares regression techniques (following Schurer et al 2013, see also Allen and Stott, 2003). The outcome are scaling factors that determine the amplitude of the fingerprints of the forcing response in the reconstructions. A forcing is detected if a scaling value of zero can be rejected at the 5% significance level, indicating that it is unlikely that climate variability alone is responsible for the similarity between forced response and reconstruction. If the 5-95% range of scaling factors encompasses one then the magnitude of the response to forcing is found to be consistent in simulations and in the reconstruction (see supplement sect. S2.4).

Figure 10 shows the results of the detection and attribution analysis using the multi-model ensemble mean, which is calculated as the mean of all model simulations described in section 2.2, except for the high-solar COSMOS and CESM1 simulations as they include a clearly different forcing. All reconstructions and models used were first filtered using a 23-year Hamming window.

The response to external forcing is detectable ($p<0.05$) in all four reconstructions from the Northern Hemisphere and during all time periods (scaling range always greater than zero; indicating that the level of agreement between the multimodel mean and the reconstructions exceeds that from random control samples significantly). The scaling ranges always encompass the scaling factor 1, which shows that the model results are consistent with the reconstructions because they do not need to be scaled up or down.

The only exceptions are the earliest time period (864-1350) for North America (tree-ring based reconstruction) where only 150 years of data were available and the early European period, which fails the residual consistency check, indicating that the residual that is attributed to internal variability is larger than expected from model simulations, possibly due to non-climatic noise in reconstructions. The results for the latter case suggest that the basic hypotheses underlying the methodology are violated because the model-reconstruction discrepancy cannot be explained by internal variability alone. External forcing is also detectable when the model and reconstruction data from all Northern Hemisphere regions are combined.

External forcing is not detectable in South America (no scaling ranges significantly larger than 0) and only for certain time periods for Antarctica and Australasia (with fits for Australasia failing the residual consistency check). External forcing is also not detectable in the combined Southern Hemisphere reconstruction. As well as being un-detectable, despite accounting for uncertainty in simulated signals due to variability, the estimated signals are also significantly smaller than simulated. Consequently, the models appear to simulate a magnitude and pattern of external forcing in the Southern Hemisphere significantly different from that derived from the PAGES 2k reconstructions. This could be due to strong noise in reconstructions swamping the forced response, calibration uncertainty in reconstructions misestimating the magnitude of the forced response, or errors in climate models as discussed below

## 6. Discussion

In the light of the results presented in the Sections 3 to 5, we discuss below each of the three questions raised in the introduction.

### 6.1 Are the statistical properties of surface temperature data for each individual continent-scale region consistent between simulations and reconstructions?

The analyses herein show that the answer to this question depends on the specific feature assessed. The simulation results and reconstructions agree at regional scale for some metrics, but disagree in many cases. The consistency between simulations and observations is still generally more robust at hemispheric and global scales, and the fit to reconstructions is improved for ensemble mean of simulations compared to individual members.

Overall, smoothed simulated temperature anomalies from the long-term average lie within the range of the originally published uncertainty estimates of the reconstructions. However, these uncertainty ranges are, for all regions except North America, defined for data at annual resolution and therefore are very likely larger than uncertainty ranges adapted for the smoothed versions of the data (see section 3.1). Thus, the published uncertainties for the reconstructions are in most cases too large to provide strong constraints on the ensemble of simulations, as different forcing amplitudes and responses are nevertheless consistent within the range of the reconstructed values. Some common signals between model results and reconstructions can be identified visually as isolated events, such as the cooling during the early 19$^{th}$ century in many regions, but they are relatively rare.

The time series for forced simulations are nevertheless significantly correlated with temperature reconstructions, for many regions, when the entire series are considered (Supplementary text S3). Models also have some skill compared to a simple *a priori* estimate assuming no temperature change over the past millennium (section 3.4). Despite using a very simple reference method as a benchmark, however, such skill is achieved nearly exclusively for Northern Hemisphere regions, specifically for the Arctic and in some models for Europe and Asia. This is in agreement with the conclusions derived from the application of the Sundberg et al. (2012) evaluation framework (section 5.2) that forced simulations are significantly closer to the reconstructions than unforced simulations in Northern Hemisphere regions. In particular, the Arctic region shows a robust agreement, as do Europe and Asia to a lesser degree. In contrast, for all the regions of the Southern Hemisphere, the models have nearly no skill compared to a constant climate reference and individual forced simulations are in most cases not significantly closer to reconstructions than an unforced reference.

The diagnostics mentioned above addressed whether simulated time series of surface temperature at continental scale have temporal similarities with reconstructed ones. The climatological or probabilistic consistency is complementary as it focuses on the distribution of temperature data, independent of the particular trajectory over time. For most regions, no consistency is found between the distribution of model results and reconstructed temperatures when using the original reconstruction uncertainty estimates (section 3.3), which are annually resolved in all cases except North America.

It should be noted, however, these results depend strongly on the uncertainty estimates considered (Bothe et al., 2013a,b): the greater the assumed reconstruction uncertainty, the weaker the consistency with model simulations as the models tend to appear over-dispersive. When reducing the uncertainties, to adapt them to the smoothing or temporal averaging applied here, the consistency improves in many regions. Such reduction of the uncertainties may, however, lead to overconfident conclusions if the original uncertainty estimates at the annual resolution did not account for all existing sources of uncertainty.

A visual comparison suggests that the temperature reconstructions show slightly more centennial to multi-centennial variability over the full period with stronger long-term trends, while model results indicate a stronger recent warming compared to some of the reconstructions (section 3.1). Comparison of the series spectra (section 3.2) reveals marked differences between regions in how well the simulations agree with the reconstructions. The best overall agreement is seen for the Arctic, where the model spectra mostly lie within a 95% confidence interval for the reconstruction spectrum. For all other regions, the model spectra often lie outside the confidence interval for some frequency ranges. The mismatch is most pronounced for South America, but there are other examples with both lower and higher variance at different frequencies in model results compared to reconstructions.

The disagreements can have various origins, in either reconstructions or simulations or both. For instance, the total variance of reconstructions is dependent on how they were calibrated to instrumental observations (e.g., Kutzbach et al., 2011) but the shape and slopes of their spectra are determined by spectra of both the true climate and the non-climatic proxy-data noise and by the signal-to-noise ratio (Moberg et al., 2008). Some studies have suggested that reconstruction methodologies may alone underestimate low-frequency variability, in addition to any frequency biases inherent to the proxy data (e.g., Smerdon et al., 2010; Esper et al., 2012; Smerdon et al., 2015). The amplitude of the reconstructed past forcing changes, which affect the model spectra, is still uncertain (Schmidt et al., 2011, 2012). The modelled transient climate response and the amplitude of internal variability at the regional scale vary considerably and thus deficiencies in applied forcings or internal model physics can lead to errors in the modelled spectra. Nevertheless, no major, systematic model underestimation of low frequency variability can be deduced at the continental scale from the analyses performed herein, in contrast to some recent studies devoted to the ocean surface temperature (Laepple and Huybers, 2014ab).

### 6.2 Are the cross regional relations of temperature variations similar in reconstructions and models?

Discrepancies in the interregional relations between reconstructions and model results are clearer than for each individual region. While the strong correlations between the temperature variations in regions from the Northern Hemisphere in model simulations have some similarities to the ones in the reconstructions (section 4.1), the correlation between the hemispheres and between the Southern Hemisphere regions are much stronger in models than in reconstructions, as previously reported by Neukom et al. (2014) at hemispheric scale.

This result is robust as it is also reflected in the larger variance explained by the first EOF mode in models than in the temperature reconstructions (section 4.2) and this is valid for most of the past millennium (section 4.3). These differences may be due to a stronger response to forcings in the models, to unrealistic internal variability in the models, or to non-climatic noise in the proxies or due to a combination of these factors, as discussed in more detail below. Additionally, there are large differences between the various models in the Southern Hemisphere. For instance Antarctic temperature is strongly related to other regional temperatures in some simulations and not in others, suggesting that specific model dynamics may account for some of the discrepancies with the reconstructions.

### 6.3 Can the signal of the response to external forcing be detected on continental scale and, if so, how large are these signals?

The agreements or disagreements between model results and reconstructions can be partly explained by the model response to forcing. The contribution of the forcing derived from simulated results can be detected in the reconstructions for all regions of the Northern Hemisphere (section 5.3). The forcings used in the PMIP3 model experiment result in simulated temperature histories that, on the whole, explain a significant fraction of the past regional temperatures in the pre-industrial climate.

This strongly contributes to the model skill for the Northern Hemisphere, as unforced internal stochastic variability is unlikely to agree between model results and observations. This is confirmed by the significant correlation coefficients (Fig. S1) and correlation statistics ($U_R$) (section 5.2) that indicate common external forcing variations. Furthermore, the correlations increase for the ensemble average relative to the available single-model simulations due to the fact that the contributions from internal variability are reduced by averaging.

On interannual time scales, the model response to volcanic forcing appears larger than represented in the reconstructions (section 5.1). There is some debate on the potential underestimation or overestimation of the cooling due to volcanic eruptions in reconstructions (e.g., Mann et al., 2012; Anchukaitis et al., 2012; Tingley et al., 2014; Büntgen et al., 2015). Nevertheless, this model overestimation was also found when compared to instrumental data and at hemispheric scale, suggesting a robust phenomenon (Brohan et al., 2012; Fernandez-Donado et al., 2013; Masson-Delmotte et al., 2013; Schurer et al., 2013). Both model results and reconstructions also show that volcanic activity impacts temperature at multidecadal timescales, with a similar magnitude of the temperature response in models and reconstructions over most of the regions in the Northern Hemisphere. This is consistent with the detection and attribution analysis (section 5.3), which indicates that the magnitude of the simulated response to forcing in the Northern Hemisphere has the correct amplitude for smoothed time series.

The role of solar forcing is less clear and none of the pre-PMIP3 COSMOS simulations with either a moderate or a weak solar forcing gives a systematically better agreement with the reconstructions than the other, although the ensemble with low solar forcing yield a better fit during the first 500 years (Fig. S18). This confirms earlier results obtained at the hemispheric scale (Masson-Delmotte et al., 2013; Schurer et al., 2014).

In the Southern Hemisphere, the influence of external forcing is often not detected (section 5.3). This is consistent with the lower correlation coefficients (Fig. S1) and weaker correlation statistics ($U_R$) there (section 5.2). The models also seem to overestimate the response compared to the signal recorded in the Southern Hemisphere reconstructions (section 5.2-5.3). This finding is likely related to the larger covariability seen within Southern Hemisphere regions in models compared to reconstructions. Moreover, control simulations display low correlations between the Northern and Southern Hemisphere regions.

The analysis performed herein, however, cannot reveal the origin of the mismatch between simulation results and reconstructions. These differences may be due to biases in the dynamics of the climate models or to errors in the implemented forcing, in particular in their spatial distribution. Land-use changes, which are not included in some models (Table 1), tend to reduce the spatial correlation between regions as deforestation did not occur at the same time over all continents (Pongratz et al., 2008; Kaplan et al., 2011).

The spatial distribution of volcanic aerosols may also contribute to pronounced regional differences. Volcanic forcing is usually not implemented as a direct simulation of changes in stratospheric sulphate concentrations due to individual eruptions, but as a mean change in the optical depth for different latitudinal bands. This can have an impact on the overestimation of the response in individual simulations or to individual eruptions. Additionally, if the latitudinal distribution of volcanic aerosols is too homogeneous, thereby inducing unrealistically symmetric forcing between hemispheres, it would also overestimate the global signature of the induced cooling.

Any non-climatic noise in the reconstruction would tend to reduce the covariance in reconstructions compared to model results, which would lead to an underestimation of the relative contribution of the forced signal. Despite the large progress made over the last few years, this may still be a critical problem in the Southern Hemisphere, where fewer long paleoclimate records are available compared to the Northern Hemisphere, explaining some of the model-data mismatch there.

The role of internal variability in driving temperature variations may also be underestimated in model simulations, particularly in the ocean dominated Southern Hemisphere, as suggested by Neukom et al. (2014). Simulated internal variability may, however, be overestimated, as reported here in at least one model and elsewhere for ENSO-type variability (Jungclaus et al., 2006) or for the Southern Ocean ice extent (Zunz et al., 2013). This would imply a ratio between internal and forced variability that is incorrect, which would lead to biased correlations between the different regions.

Another potential explanation for the differences in the spatial covariance structure between models and observations relates to the relatively coarse resolution of the climate models. Using models with higher spatial resolution will increase the number of spatial degrees of freedom and potentially improve the co-variance structure of the climate models compared to reconstructions. Nevertheless, the expense required for both high spatial and temporal resolution, as well as the necessary ensemble approach could be prohibitive.

## 7. Conclusions and perspectives

The analysis of model simulations and PAGES 2k temperature reconstructions has allowed us to extend some of the some conclusions previously articulated at only hemispheric scale. For all the continental-scale regions in the Northern Hemisphere, the models are able to simulate a forced response with a magnitude similar to the one derived from reconstructions. Despite higher levels of variability on continental scales (relative to full hemispheres), the role of forcing is found to be important. This leads to reasonable agreement between models and temperature reconstructions.

Nevertheless, a deeper assessment of the consistency between simulated results and reconstruction is limited because of the large uncertainties in the reconstructions and the weak constraints on the estimates of this uncertainty. Notably, the agreement between simulation results and reconstructions is poor for the Southern Hemisphere regions. Our results indicate that models have a much clearer response to forcing than deduced from the reconstructions, leading to a greater consistency across the Southern Hemisphere regions and between hemispheres in model results than in the reconstructions.

It is not possible to precisely assess which part of those disagreements comes from the biases in model dynamics, the forcing or in the reconstructions. As suggested in many previous studies, substantial progress will only be possible with better uncertainty quantification and reduction (spatially and temporally) in the reconstructions and the forcing, and through model improvements.

Nevertheless, on the basis of our results we highlight four specific points that may lead to significant advances in the coming years.

The first is the insights that can be gained through studying the discrepancies between reconstructions and simulations relative to direct observations over the most recent decades. A quantitative comparison between simulations, reconstructions, and instrumental data would provide useful information on the origin of those disagreements, allow an estimate of the non-climatic noise in reconstructions, and would elucidate how mismatches over the last 150 years are related to disagreements over the last several millennia (e.g., Ding et al 2014).

Secondly, large uncertainties are associated with the behaviour of the ocean over the past millennium. The discrepancies in the low frequency variability between model results and reconstructions at the continental scale seem less systematic than for some oceanic data (Laepple and Huybers, 2014ab), but clearly assessing this would require additional analyses. As new paleoclimate data compilations are now available for the global ocean (Tierney et al., 2015; McGregor et al., 2015), model-data comparison for oceanic regions should be encouraged, and the compatibility between ocean and land temperature reconstructions tested. This would allow us to assess the multidecadal internal and forced variability of the ocean and to determine if it is the origin of the disagreement between model simulations and Southern Hemisphere reconstructions (e.g., Neukom et al., 2014). Internal ocean variability can also have a significant influence on Northern Hemisphere climate as seen in several studies investigating the circulation in the Atlantic at multi-decadal time scales (e.g., Delworth and

Mann, 2000; Knight et al., 2005; Lohmann et al., 2014). These are the timescales for which most models tend to display less variability than reconstructions.

Third, our comparison of continental-scale temperature reconstructions with simulated temperatures only uses a small fraction of the information provided by models and paleoclimate records. As discussed in Phipps et al. (2013), other approaches can provide analyses complementary to classical model-data comparison, through a better handling of the various sources of uncertainty. Promising examples are proxy forward models, which simulate directly the proxy records from climate model outputs (e.g., Evans et al., 2013) and data assimilation methods (e.g., Widmann et al., 2010; Goosse et al., 2012b; Steiger et al., 2014). These approaches combine model results and observations to obtain the best estimates of past change and may be most effective at to detecting inconsistencies between model and palaeoclimate estimates.

Finally, one could also question the selection of the continental scale as the basis of a comparison, as regional changes are strongly affected by modes of variability such as ENSO, the Southern Annular Mode, the North Atlantic Oscillation or the Pacific North America pattern. These modes could imprint temperature patterns that are masked by averaging over the continents. On the other hand, model-data comparison made at smaller spatial scales has revealed highly variable and even contradictory results at nearby regions (Moberg et al., 2015), suggesting that a large number of local proxy data sites are needed for obtaining robust results. Ideally, a sub-regional selection from key teleconnection regions should be used to assess the stability of climate modes (Raible et al., 2014) or enable reliable reconstruction of modes of variability (Lehner et al., 2012; Zanchettin et al., 2015; Ortega et al., 2015), although this requires strong reconstruction skill to be successful (e.g., Russon et al., 2015). Additionally, spatially resolved reconstructions should be targeted because they offer useful potential for dynamic interpretation (e.g., Luterbacher et al., 2004; Mann et al., 2009; Steiger et al., 2014, PAGES 2k Consortium, 2014; Shi et al., 2015).

In summary, our results for the Northern Hemisphere suggest a convergence of our understanding of climate variability over the past 1000 years, but there remain many open questions for the Southern Hemisphere. Progress may be expected from comparing simulations, reconstructions and observations in the instrumental period, from a better knowledge of internal and forced variability of the ocean, from efforts to understand climate variability via proxy forward modelling and data assimilation, and from a clearer view of the influence of climate modes on temperature variability.

## Acknowledgements

This study is based on discussions held during the joint workshop of the PAGES 2k network and PAST2k-PMIP *Integrated analyses of reconstructions and multi-model simulations for the past two millennia*, Madrid, Spain, 4-6 November 2013. PAGES and FECYT (FCT-13-6276) are greatly thanked for supporting this workshop. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP. The U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support for CMIP and led the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. H.G. is Research Director with the Fonds National de la Recherche Scientifique (F.R.S.-FNRS-Belgium). This work is supported by the F.R.S.- FNRS and by the Belgian Federal

Science Policy Office (Research Program on Science for a Sustainable Development). CCR and FL are supported by the Swiss National Science foundation. PY is supported by the MILEX project of the Swedish Research Council. JG is funded by Australian Research Council Project DE130100668. OB was supported by LOCHMES (Leibniz-Society), PRIME-II (within DFG INTERDYNAMIK) and CliSAP. LFD was funded by a FPU grant: AP2009-4061. AM and AH are supported by the Swedish Research Council grants B0334901 and C0592401. GH and AS are supported by the ERC advanced grant TITAN (320691).

**Author contributions**

LFD, FGR, EGB, HG, JJ organized the workshop at the origin of this paper. HG led the synthesis. OB, HG, GH, AM, CR, AS, SW, EZ coordinated the writing. LFD, EGB AH, FL, NM prepare the data sets and made them available to the whole group. OB, LFD, EGB, FGR, AH, FL, NM, AM, AS, SW, EW, MW, EZ performed the figures and their initial analysis. All authors contribute to the writing of the various sections and reviewed the manuscript.

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, Nat. Climate Change, 4, 564–569, 2014.

Allen, M. R., and P. A. Stott: Estimating signal amplitudes in optimal fingerprinting, Part I: Theory, Clim. Dyn. 21, 477-491, 2003.

Ammann, C. M., Joos, F., Schimel, D. S., Otto-Bliesner, B. L., and Tomas, R. A.: Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model. Proc. Nat. Acad. Sciences, 104(10), 3713-3718. Doi: 10.1073/pnas.0605064103, 2007.

Anchukaitis, K.J., Breitenmoser, P., Briffa, K. R., Buchwal, A., Büntgen, U., Cook, E. R., D'Arrigo, R. D., Esper, J., Evans, M. N., Frank, D., Grudd, H., Gunnarson, B. E., Hughes, M. K., Kirdyanov, A. V., Körner,C., Krusic, P. J., Luckman, B., Melvin, T. M., Salzer, M. W., Shashkin, A.V., Timmreck, C., Vaganov, E. A., and Wilson, R. J. S.: Tree rings and volcanic cooling, Nat. Geosci., 5(12), 836–837, doi:10.1038/ngeo1645, 2012.

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996

Annan, J. D., and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, doi:10.1029/2009GL041994, 2010.

Ault, T. R., Deser, C., Newman, M. and Emile-Geay, J.: Characterizing decadal to centennial variability in the equatorial Pacific during the last millennium. Geophys. Res. Let. 40 (13), 3450-3456, doi: 10.1002/grl.50647, 2013.

Bard, E., Raisbeck, G., Yiou, F., and Jouzel, J.: Solar irradiance during the last 1200 years based on cosmogenic nuclides. Tellus B, 52(3), 985-992. Doi: 10.1034/j.1600-0889.2000.d01-7.x, 2000.

Berger, A.: Long-term variations of daily insolation and Quaternary climatic changes, J. Atmos. Sciences, 35(12), 2362-2367. Doi: 10.1175/1520-0469(1978)035<2362:LTVODI>2.0.CO;2, 1978.

Bertrand, C., Loutre, M.-F., Crucifix, M., and Berger, A.: Climate of the last millennium: a sensitivity study. Tellus 54A, 221-244, 2002.

Bindoff, N.L., Stott, P.A., AchutaRao, K.M., Allen, M.R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I.I., Overland, J., Perlwitz, J., Sebbari, R., and Zhang, X. : Detection and Attribution of Climate Change: from Global to Regional. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2013.

Bloomfield, P.: Fourier Analysis of Time Series: An Introduction. New York: John Wiley and Sons, 1976.

Borlace, S., Cai, W. and Santoso, A.: Multidecadal ENSO Amplitude Variability in a 1000-yr Simulation of a Coupled Global Climate Model: Implications for Observed ENSO Variability, J. Clim. 26 (23), 9399-9407, 2013.

Bothe, O., Jungclaus, J. H., Zanchettin, D., and Zorita, E.: Climate of the last millennium: ensemble consistency of simulations and reconstructions, Clim. Past, 9, 1089-1110, doi:10.5194/cp-9-1089-2013, 2013a.

Bothe, O., Jungclaus, J. H., and Zanchettin, D.: Consistency of the multi-model CMIP5/PMIP3-past1000 ensemble, Clim. Past, 9, 2471-2487, doi:10.5194/cp-9-2471-2013, 2013b.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe Ouchi, A., Otto-Bliesner, B., and Zhao, Y.,: Evaluation of climate models using paleoclimate data, Nature Climate Change, 2, 417–424 doi:10.1038/NCLIMATE1456, 2012.

Bretagnon, P., and Francou, G.: Planetary theories in rectangular and spherical variables-VSOP 87 solutions. Astronomy and Astrophysics, 202, 309-315, 1988.

Brohan, P., Allan, R., Freeman, E., Wheeler, D., Wilkinson, C., Williamson, F.: Constraining the temperature history of the past millennium using early instrumental observations, Clim. Past, 8, 1551-1563, doi:10.5194/cp-8-1551-2012, 2012

Büntgen, U., Trnka, M., Krusic, P. J., Kyncl, T., Kyncl, J., Luterbacher, J., Zorita, E., Charpentier Ljungqvist, F., Auer, I., Konter, O., Schneider, L., Tegel, W., Štěpánek, P., Brönnimann, S., Hellmann, L., Nievergelt, D., and Esper, J.: Tree-Ring Amplification of the Early-19th Century Summer Cooling in Central Europe, J. Clim., in press, doi: http://dx.doi.org/10.1175/JCLI-D-14-00673.1, 2015

Christiansen, B., Schmith, T., and Thejll, P.: A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness, J. Clim., 22, 951–976, doi:10.1175/2008JCLI2301.1, 2009.

Coats, S., Cook, B.I., Smerdon, J.E., and Seager, R.: North American Pan-Continental droughts in model Simulations of the last millennium, J. Clim 28, 2025-2043, 2015a.

Coats, S., Smerdon, J.E., Cook, B.I., and R. Seager, R.: Are simulated megadroughts in the North American Southwest forced?, J. Clim., 28, 124-142. doi:http://dx.doi.org/10.1175/JCLI-D-14-00071, 2015b.

Comboul, M., Emile-Geay, J., Hakim, G. J. and Evans, M. N.: Paleoclimatic sampling as a sensor placement problem. J. Clim.. 28, 7717-7740, 2015

Cook, B.I., Ault, T.R., and Smerdon, J.E.: Unprecedented 21st-century drought risk in the American Southwest and Central Plains, Science Advances, 1, e1400082, 2015.

Cook, E.R., Meko, D. M., Stahle, D. W., and Cleaveland, M. K: Drought reconstructions for the continental United States. J. Climate, 12, 1145–1162. doi: http://dx.doi.org/10.1175/1520-0442(1999)012<1145:DRFTCU>2.0.CO;2., 1999.

Cook, E.R., Woodhouse, C. A., Eakin, C. M., Meko, D. M., Stahle, D. W.: Long-term aridity changes in the Western United States, Science 306, 1015, doi: 10.1126/science.1102586, 2004.

Cook, E.R., Anchukaitis, K. J., Buckley, B. M., D'Arrigo, R. D., Jacoby, G. C., Wright, W. E.: Asian monsoon failure and megadrought during the Last Millennium, Science, 328, 486, 2010.

Crowley, T.J.: Causes of Climate Change Over the Past 1000 Years, Science, 289, 270-277, 2000.

Crowley, T.J. and Unterman, M. B.: Technical details concerning development of a 1200-yr proxy index for global volcanism, Earth System Science Data, Vol. 5, pp. 187-197, DOI: 10.5194/essd-5-187-2013, 2013.

D'Arrigo, R., Wilson, R., Jacoby, G., On the long-term context for late twentieth century warming, J. Geophys. Res., 111, D03103, doi:10.1029/2005JD006352, 2006.

Delworth, T.L., and Mann, M.E.: Observed and simulated multidecadal variability in the Northern Hemisphere, Clim. Dyn 16:661-676, 2000.

Ding, Y., Carton, J. A., Chepurin, G. A., Stenchikov, G., Robock, A., Sentman, L. T., and Krasting, J. P.: Ocean response to volcanic eruptions in Coupled Model Intercomparison Project 5 simulations, J. Geophys. Res. Oceans, 119, 5622–5637, doi:10.1002/2013JC009780, 2014.

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., de Noblet, N., Duvel, J.-P., Ethé, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J.-Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M.-P., Lefevre, F., Levy, C., Li, Z. X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, and N., Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim. Dyn. 40(9-10), 2123-2165, 2013.

Emile-Geay, J., Cobb, K. M., Mann, M. E., and Wittenberg, A. T.: Estimating Central Equatorial Pacific SST Variability over the Past Millennium. Part I: Methodology and Validation, J. Climate, 26, 2302–2328, doi: http://dx.doi.org/10.1175/JCLI-D-11-00510.1, 2013

Esper, J., Frank, D., Wilson, R., and Briffa, K.: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium, Geophys. Res. Lett., 32, L07711, doi:10.1029/2004GL021236, 2005.

Esper, J., Frank, D.C., Timonen, M., Zorita, E., Wilson, R. J. S., Luterbacher, J., Holzkämper, S., Fischer, N., Wagner, S., Nievergelt, D., Verstege, A, and Büntgen, U.: Orbital forcing of tree-ring data, Nature Clim. Change, 2, 862–866, 2012.

Evans, M.N., Tolwinski-Ward, S.E., Thompson, D.M., and Anchukaitis, K.J.: Applications of proxy system modeling in high resolution paleoclimatology, Quate. Science Rev. 76, 16-28, 2013.

Fernández-Donado, L., González-Rouco, J. F., Raible, C. C., Ammann, C. M., Barriopedro, D., García-Bustamante, E., Jungclaus, J. H., Lorenz, S. J., Luterbacher, J., Phipps, S. J., Servonnat, J., Swingedouw, D., Tett, S. F. B., Wagner, S., Yiou, P., and Zorita, E.: Large-scale t temperature response to external forcing in simulations and reconstructions of the last millennium, Clim. Past, 9, 393-421, DOI: doi:10.5194/cp-9-393-2013, 2013.

Feulner, G.: Are the most recent estimates for Maunder Minimum solar irradiance in agreement with temperature reconstructions? Geophysical Research Letters, 38, L16706, doi:10.1029/2011GL048529, 2011.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. and Rummukainen, M.: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Flückiger, J., Dällenbach, A., Blunier, T., Stauffer, B., Stocker, T. F., Raynaud, D., and Barnola, J. M.: Variations in atmospheric N2O concentration during abrupt climatic changes. Science, 285(5425), 227-230. doi: 10.1126/science.285.5425.227, 1999.

Flückiger, J., Monnin, E., Stauffer, B., Schwander, J., Stocker, T. F., Chappellaz, J., and Barnola, J. M.: High resolution Holocene N2O ice core record and its relationship with CH4 and CO2, Glob. Biogeochem. Cycles, 16(1), 10-1, doi: 10.1029/2001GB001417, 2002

Franke, J., Frank, D., Raible, C. C., Esper, J. and Brönnimann, S.: Spectral biases in tree-ring climate proxies, Nat. Clim. Change, 3, 360–364, doi:10.1038/nclimate1816, 2013.

Gao, C., Robock, A., and Ammann, C.: Volcanic forcing of climate over the last 1500 years: An improved ice-core based index for climate models, J. Geophys. Res., 113, D2311, doi:10.1029/2008JD010239, 2008.

Gergis, J., Neukom, R., Gallant, A. J. E. and Karoly, D. J.: Evidence of rapid late 20th century warming from Australasian temperature reconstruction ensembles spanning the last millennium, J. Clim. (submitted), 2015.

Gonzalez-Rouco, J.F., Beltrami, H., Zorita, E., and von Storch, H.: Simulation and inversion of borehole temperature profiles in surrogate climates: spatial distribution and surface coupling, Geophys. Res. Let. 33, L01703, doi:10.1029/2005GL024693, 2006.

Goosse, H., Masson-Delmotte, V., Renssen, H., Delmotte, M., Fichefet, T., Morgan, V., van Ommen, T., Khim, B. K., and Stenni, B.: A late medieval warm period in the Southern Ocean as a delayed response to external forcing? Geophysical Research Letters 31 (6): L06203, 2004.

Goosse, H., Renssen, H., Timmermann, A., and Bradley, R.S.: Internal and forced climate variability during the last millennium: a model-data comparison using ensemble simulations, Quat. Science Rev., 24, 1345-1360, 2005.

Goosse H., Arzel, O., Luterbacher, J., Mann, M. E., Renssen, H., Riedwyl, N., Timmermann, A., Xoplaki, E., and Wanner, H.. The origin of the European "Medieval Warm Period". Clim. Past 2, 99-113, 2006.

Goosse, H., Braida, M., Crosta, X., Mairesse, A., Masson-Delmotte, V., Mathiot, P., Neukom, R., Oerter, H., Philippon, G., Renssen, H., Stenni, B., van Ommen, T., and Verleyen, E.: Antarctic temperature changes during the last millennium: evaluation of simulations and reconstructions. Quaternary Science Reviews 55: 75-90, 2012a.

Goosse H., Crespin, E, Dubinkina, S., Loutre, M.F., Mann, M. E, Renssen, H., Sallaz-Damaz, Y., and Shindell, D.: The role of forcing and internal dynamics in explaining the “Medieval Climate Anomaly”, Clim. Dyn. 39, 2847–2866, doi: 10.1007/s00382-012-1297-0, 2012b.

Hansen, J., and Sato, M.: Greenhouse gas growth rates, Proc. Natl. Acad. Sci., 101, 16109-16114, doi:10.1073/pnas.0406982101, 2004.

Hargreaves, J. C., Paul, A., Ohgaito, R., Abe-Ouchi, A., and Annan, J. D.: Are paleoclimate model ensembles consistent with the MARGO data synthesis? Climate of the Past 7, 917–933, doi:10.5194/cp-7-917-2011, 2011.

Hargreaves, J. C., Annan, J. D., Ohgaito, R., Paul, A., and Abe-Ouchi, A.: Skill and reliability of climate model ensembles at the Last Glacial Maximum and mid-Holocene. Clim. Past, 9, 811-823, doi:10.5194/cp-9-811-2013, 2013.

Hegerl, G. C., Crowley, T. J., Baum, S. K., Kim, K.-Y., and Hyde, W. T.: Detection of volcanic, solar and greenhouse gas signals in paleo-reconstructions of Northern Hemispheric temperature, Geophys. Res. Lett., 30, 1242, doi:10.1029/2002GL016635, 5, 2003.

Hegerl, G. C., Crowley, T. J., Hyde, W. T. and Frame, D. J.: Climate sensitivity constrained by temperature reconstructions over the past seven centuries, Nature 440: 1029-1032, 2006.

Hegerl, G. C., Crowley, T. J., Allen, M., Hyde, W. T., Pollack, H. N., Smerdon, J., and Zorita, E.: Detection of human influence on a new, validated 1500-year temperature reconstruction, J. Clim., 20, 650-666, doi:10.1175/JCLI4011.1, 2007.

Hegerl, G., Luterbacher, J., González-Rouco, F., Tett, S., Crowley, T., and Xoplaki, E.: Influence of human and natural forcing on European seasonal temperatures, Nat. Geoscience 4 (1179): 99–103, 2011.

Hegerl, G.C., and Zwiers, F.W.: Use of models in detection and attribution of climate change, WIRES: Climate Change, 2, 570-591, 2011.

Hind, A., and Moberg, A., and Sundberg, R.: Statistical framework forevaluation of climate model simulations by use of climate proxy data from the last millennium – Part 2: A pseudo-proxy study addressing the amplitude of solar forcing, Clim. Past, 8, 1355– 1365, doi:10.5194/cp-8-1355-2012, 2012.

Hind, A., and Moberg, A. : Past millennial solar forcing magnitude. A statistical hemispheric-scale climate model versus proxy data comparison, Clim. Dyn., 41, 2527–2537, doi:10.1007/s00382-012-1526-6, 2013.

Hurtt, G. C., Chini, L. P., Frolking, S., Betts, R. A., Feddema, J., Fischer, G., and Wang, Y. P.: Harmonization of land-use scenarios for the period 1500–2100: 600 years of global gridded annual land-use transitions, wood harvest, and resulting secondary lands, Clim. Change, 109(1-2), 117-161. Doi: 10.1007/s10584-011-0153-2, 2011

Jansen, E., Overpeck, J., Briffa, K.R., Duplessy, J.-C., Joos, F., Masson-Delmotte, V., Olago, D., Otto-Bliesner, B., Peltier, W.R., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R., and Zhang, D.: Palaeoclimate. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Johns, T. C., Gregory, J. M., Ingram, W. J., Johnson, C. E., Jones, A., Lowe, J. A., and Woodage, M. J.: Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios, Clim. Dyn., 20(6), 583-612, doi: 10.1007/s00382-002-0296-y, 2003

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

Jones, P.D. and Mann, M.E.: Climate over past millennia, Rev. Geophys. 42, RG2002, doi: 10.1029/2003RG000143, 2004.

Jungclaus, J. H., Keenlyside, N., Botzet, M., Haak, H., Luo, J.-J., Latif, M., Marotzke, J., Mikolajewicz, U., and Roeckner, E.: Ocean Circulation and Tropical Variability in the Coupled Model ECHAM5/MPI-OM, J. Clim., 19, 3952– 3972, 2006.

Jungclaus, J. H., Lorenz, S. J., Timmreck, C., Reick, C. H., Brovkin, V., Six, K., and Marotzke, J.: Climate and carbon-cycle variability over the last millennium. Climate of the Past, 6, 723-737. doi: 10.5194/cp-6-723-2010, 2010.

Jungclaus, J. H., Lohmann, K., and Zanchettin, D. Enhanced 20th-century heat transfer to the Arctic simulated in the context of climate variations over the last millennium. Clim. Past, 10(6), 2201-2213, doi: 10.5194/cp-10-2201-2014, 2014.

Kaplan, J. O., Krumhardt, K. M., Ellis, E. C., Ruddiman, W. F., Lemmen, C., & Goldewijk, K. K.: Holocene carbon emissions as a result of anthropogenic land cover change. The Holocene, 21(5) 775,–79, doi: 10.1177/0959683610386983, 2011.

Knight, J.R., Allan R.J., Folland C.K., Vellinga M., and Mann M.E.:, A signature of persistent natural thermohaline circulation cycles in observed climate. Geophys. Res. Let. 32 (20), L20708, 2005.

Krivova, N. A., Balmaceda, L., and  Solanki, S. K.: Reconstruction of solar total irradiance since 1700 from the surface magnetic flux, Astron. Astrophys. 467(1), 335-346, doi: 10.1051/0004-6361:20066725, 2007

Kutzbach, L., Thees, B., and Wilmking, M.: Identification of linear relationships from noisy data using errors-in-variables models-relevance for reconstruction of past climate from tree-ring and other proxy information, Clim. Change, 105, 155–177, doi 10.1007/s10584-010-9877-7, 2011.

Laepple, T., and Huybers, P.: Global and regional variability in marine surface temperatures, Geophys. Res. Lett., 41, doi:10.1002/2014GL059345, 2014a.

Laepple, T., and Huybers, P.: Ocean surface temperature variability: Large model–data differences at decadal and longer periods, Proceed. Nat. Acad. Sciences 11 (47) 16682–16687, doi: 10.1073/pnas.1412077111, 2014b.

Lamarque, J. F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., and Van Vuuren, D. P.: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application., Atmos. Chem. Phys., 10(15), 7017-7039 doi: 10.5194/acp-10-7017-2010, 2010

Landrum, L., Otto-Bliesner, B. L., Wahl, E. R., Conley, A., Lawrence, P. J., Rosenbloom, N., and Teng, H.: Last millennium climate and its variability in CCSM4, J. Clim. 26(4), 1085-1111, doi: 10.1175/JCLI-D-11-00326.1, 2013

Lehner, F., Raible, C. C., and Stocker, T. F.: Testing the robustness of a precipitation proxy-based North Atlantic Oscillation reconstruction, Quat. Sci. Rev., 45, 85-94, 2012.

Lehner, F., Born, A., Raible, C. C., Stocker, T. F.: Amplified inception of European Little Ice Age by sea ice-ocean-atmosphere feedbacks, J. Clim. 26, 7586-7602, doi: 10.1175/JCLI-D-12-00690.1, 2013.

Lehner, F., Joos, F., Raible, C. C., Mignot, J., Born, A., Keller, K. M., and Stocker, T. F.: Climate and carbon cycle dynamics in a CESM simulation from 850-2100 CE, Earth Sys. Dyn. Discuss., 6, 1–56, 2015.

Lohmann, K., Jungclaus, J. H., Matei, D., Mignot, J., Menary, M., Langehaug, H. R., Ba, J., Gao, Y., Otterå, O. H., Park, W., and Lorenz, S.: The role of subpolar deep water formation and Nordic Seas overflows in simulated multidecadal variability of the Atlantic meridional overturning circulation, Ocean Sci., 10, 227-241, doi:10.5194/os-10-227-2014, 2014

Lorenz, E.: Deterministic Nonperiodic Flow, J. Atmos. Sci., 20, 130–141, 1963.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European seasonal and annual temperature variability, trends, and extremes xince 1500, Science 5: 303, 1499-1503, doi:10.1126/science.1093877, 2004.

MacFarling Meure, C., Etheridge, D., Trudinger, C., Steele, P., Langenfelds, R., Van Ommen, T., and .Elkins, J.: Law Dome CO2, CH4 and N2O ice core records extended to 2000 years BP. Geophysical Research Letters, 33(14). Doi: 10.1029/2006GL026152, 2006.

Machida, T., Nakazawa, T., Fujii, Y., Aoki, S., and Watanabe, O.: Increase in the atmospheric nitrous oxide concentration during the last 250 years, Geophys. Res. Let., 22(21), 2921-2924. Doi: 10.1029/2001GB001417, 1995.

Mann, M.E., Bradley, R.S., and Hughes, M.K.: Northern Hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations, Geophys. Res. Let. 26, 759-762, 1999.

Mann, M. E., Fuentes, J. D. and Rutherford, S.: Underestimation of volcanic cooling in tree-ring-based reconstructions of hemispheric temperatures, Nat. Geosci., 5(3), 202–205, 2012.

Mann, M., Zhang, Z., Rutherford, S., Bradley, R., Hughes, M., Shindell, D., Ammann, C., Faluvegi, G. and Ni, F.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science 326, 1256–1260, 2009.

Marzban, C., Wang, R., Kong, F., and Leyton, S.: On the effect of correlations on rank histograms: Reliability of temperature and wind speed forecasts from finescale ensemble reforecasts, Month. Weath. Rev., 139(1), 295-310, 2011.

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., González Rouco, J. F., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X. and Timmermann, A.: Information from Paleoclimate Archives. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

McGregor H.V., Evans, M. N., Goosse, H., Leduc, G., Martrat, B., Addison, J. A., Mortyn, P. G., Oppo, D. W., Seidenkrantz, M.S., Sicre, M.-A., Phipps, S. J., Selvaraj, K., Thirumalai, K., Filipsson, H. L., and Ersek, V.: Robust global ocean cooling trend for the past two millennia. Nat. Geos., 8(9), 671-677, DOI:10.1038/NGEO2510, 2015.

McKay, N.P., and Kaufman, D.S.: An extended Arctic proxy temperature database for the past 2,000 years. Scientific Data 1:140026 doi: 10.1038/sdata.2014.26, 2014.

Miller, G., Geirsdóttir, Á., Zhong, Y., Larsen, D. J., Otto-Bliesner, B. L., Holland, M. M., Bailey, D. A., Refsnider, K. A., Lehman, S. J., Southon, J. R., Anderson, C., Björnsson, H.,and Thordarson, T.: Abrupt onset of the Little Ice Age triggered by volcanism and sustained by sea-ice/ocean feedbacks, Geophys. Res. Lett., 39, L02708, doi:10.1029/2011GL050168, 2012.

Moberg, A., Mohammad, R., and Mauritsen, T.: Analysis of the Moberg et al. (2005) hemispheric temperature reconstruction, Clim. Dyn. 31, 957-971, doi:10.1007/s00382-008-0392-8, 2008

Moberg, A., and Brattström, G.: Prediction intervals for climate reconstructions with autocorrelated noise—An analysis of ordinary least squares and measurement error methods, Palaeogeo., Palaeoclim., Palaeoecol. 308, 313–329, 2011

Moberg, A.: Comparisons of simulated and observed Northern Hemisphere temperature variations during the past millennium – selected lessons learned and problems encountered, Tellus B 65, 19921, doi:10.3402/tellusb.v65i0.19921, 2013.

Moberg, A., Sundberg, R., Grudd H., and Hind; A.: Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 3: Practical considerations, relaxed assumptions, and using tree-ring data to address the amplitude of solar forcing, Clim. Past 11, 425–448, doi:10.5194/cp-11-425-2015, 2015

Morice, C.P., Kennedy, J.J., Rayner, N.A. and Jones, P.D.,: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset, J. Geophys. Res., 117, D08101, doi:10.1029/2011JD017187, 2012.

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973.

Neukom, R., and Gergis, J.: Southern Hemisphere high-resolution palaeoclimate records of the last 2000 years, The Holocene 5: 501–524, 2012.

Neukom, R., Gergis, J., Karoly, D., Wanner, H., Curran, M., Elbert, J., González-Rouco, F., Linsley, B., Moy, A., Mundo, I., Raible, C., Steig, E., van Ommen, T., Vance, T., Villalba, R., Zinke, J. and Frank, D.: Inter-hemispheric temperature variability over the last millennium, Nat. Clim. Change 4: 362–367, 2014

Ortega, P., Lehner, F., Casado, M., Swingedouw, D., Masson-Delmotte, V., Yiou, P., Raible, C. C.: A multi-proxy model-tested NAO reconstruction for the last millennium. Nature, submitted, 2015.

PAGES 2k Consortium : Ahmed, M. , Anchukaitis, K., Asrat, A., Borgaonkar, H., Braida, M., Buckley, B. , Büntgen, U., Chase, B., Christie, D., Cook, E., Curran, M., Diaz, H. , Esper, J., Fan, Z.X., Gaire, N., Ge, Q., Gergis, J. , Gonzalez-Rouco, J.F., Goosse, H. , Grab, S., Graham, R., Graham, N., Grosjean, M., Hanhijärvi, S., Kaufman, D., Kiefer, T.,. Kimura, K., Korhola, A., Krusic, P., Lara, A., Lézine, A.M., Ljungqvist, F., Lorrey, A., Luterbacher, J., Masson-Delmotte, D. McCarroll, J. McConnell, N. McKay, M. Morales, A. Moy, R. Mulvaney, I. Mundo, V., Nakatsuka, T., Nash, D., Neukom, R., Nicholson, S., Oerter, H., Palmer, J., Phipps, S., Prieto, M., Rivera, A., Sano, M., Severi, M., Shanahan, T., Shao, X., Shi, F., Sigl, M., Smerdon, J., Solomina, O., Steig, E., Stenni, B., Thamban, M., Trouet, V., Turney, C., Umer, M., van Ommen, T., Verschuren, D., Viau, A., Villalba, R., Vinther, B., von Gunten, L., Wagner, S., Wahl, E., Wanner, H., Werner, J., White, J., Yasue, K., Zorita, E.: Continental-scale temperature variability during the last two millennia. Nature Geoscience 6, 339-346 DOI: 10.1038/NGEO1797, 2013.

PAGES 2k Consortium (Primary Authors: K. Anchukaitis, U. Büentgen, J. Emile-Geay, M. N. Evans, H. Goosse, D. Kaufman, J. Luterbacher, J. Smerdon, M. Tingley, L. von Gunten): A Community-driven framework for climate reconstructions, Eos Trans. AGU, 95(40), 361, 2014.

Percival, D.B., and Walden, A.T., Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques, Cambridge University Press, 1993.

Phipps, S. J., McGregor, H. V., Gergis, J., Gallant, A. J., Neukom, R., Stevenson, S. and Van Ommen, T. D.: Paleoclimate data–model comparison and the role of climate forcings over the past 1500 years, J. Clim., 26(18), 6915-6936, doi: 10.1175/JCLI-D-12-00108.1, 2013.

Pongratz, J., Reick, C., Raddatz, T., and Claussen, M.: A reconstruction of global agricultural areas and land cover for the last millennium, Glob. Biogeochem. Cycles, 22(3), doi: 10.1029/2007GB003153, 2008.

Pongratz, J., Reick, C. H., Raddatz, T., and Claussen, M.: Effects of anthropogenic land cover change on the carbon cycle of the last millennium, Glob. Biogeochem.l Cycles, 23(4), doi: 10.1029/2009GB003488, 2009.

Raible, C. C., Stocker, T. F., Yoshimori, M., Renold, M., Beyerle, U., Casty, C., and Luterbacher, J.: Northern Hemispheric trends of pressure indices and atmospheric circulation patterns in observations, reconstructions, and coupled GCM simulations, J. Climate, 18, 3968-3982, 2005.

Raible, C.C., C. Casty, J. Luterbacher, A. Pauling, J. Esper, D.C. Frank, U. Büntgen, A.C. Roesch, P. Tschuck, M. Wild, P-L. Vidale, C. Schär, and H. Wanner: Climate Variability - Observations, Reconstructions, and Model Simulations for the Atlantic-European and Alpine region from 1500-2100 AD, Clim. Change, 79, 9-29, 2006.

Raible, C. C., Lehner, F., González-Rouco, J. F., and Fernández-Donado, L.: Changing correlation structures of the Northern Hemisphere atmospheric circulation from 1000 to 2100 AD, Clim. Past, 10, 537-550, doi:10.5194/cp-10-537-2014, 2014.

Rougier, J., Goldstein, M., and House, L.: Second-order exchangeability analysis for multimodel ensembles. J. Am. Stat. Assoc., 108(503), 852-863, 2013.

Russon, T., Tudhope A., Hegerl, G.C., and Collins M.: Inferring changes in ENSO amplitude from proxy records, Geophys.l Res. Lett., in press, 2015.

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0), Geosci. Model Dev., 4, 33-45, doi:10.5194/gmd-4-33-2011, 2011.

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., Delaygue, G., Joos, F., Krivova, N. A., Muscheler, R., Otto-Bliesner, B. L., Pongratz, J., Shindell, D. T., Solanki, S. K., Steinhilber, F., and Vieira, L. E. A.: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.1), Geosci. Model Dev., 5, 185–191, doi:10.5194/gmd-5-185-2012, 2012.

Schmidt, G. A., Annan, J. D., Bartlein, P. J., Cook, B. I., Guilyardi, E., Hargreaves, J. C., Harrison, S. P., Kageyama, M., LeGrande, A. N., Konecky, B., Lovejoy, S., Mann, M. E., Masson-Delmotte, V., Risi, C., Thompson, D., Timmermann, A., Tremblay, L.-B., and Yiou, P.: Using palaeo-climate comparisons to constrain future projections in CMIP5, Clim. Past, 10, 221-250, doi:10.5194/cp-10-221-2014, 2014a.

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., and Zhang, J.: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. Journal of Advances in Modeling Earth Systems, 6(1), 141-184. doi: 10.1002/2013MS000265, 2014b.

Schurer A.P., Hegerl G.C., Mann M. E., Tett, S. F. B., and Phipps, S. J.: Separating forced from chaotic climate variability over the Past Millennium, J. Climate, 26, 6954–6973, 2013.

Schurer, A. P., Tett, S. F. and Hegerl, G. C.: Small influence of solar variability on climate over the past millennium. Nat. Geoscience, 7 (2), 104–108, doi:10.1038/NGEO2040, 2014.

Shapiro, A. I., Schmutz, W., Rozanov, E., Schoell, M., Haberreiter, M., Shapiro, A. V., and Nyeki, S.: A new approach to the long-term reconstruction of the solar irradiance leads to large historical solar forcing, Astron. Astrophys., 529, A67, doi:10.1051/0004-6361/201016173, 2011

Shi, F., Yang, B., Mairesse, A., von Gunten, L., Li, J., Bräuning, A., Yang, F., Xiao, X.: Northern Hemisphere temperature reconstruction during the last millennium using multiple annual proxies. Climate Research, 56(3), 231-244, 2013.

Shi, F., Ge, Q., Yang, B., Li, J., Yang, F., Charpentier Ljungqvist, F., Solomina, O., Nakatsuka, T., Wang, N., Zhao, S., Xu, C., Fang, K., Sano, M., Chu, G., Fan, Z., Gaire, N. P., Zafar, M. U.: A multi-proxy reconstruction of spatial and temporal variations in Asian summer temperatures over the last millennium. Climatic Change,131(4): 663-676, 2015.

Shindell, D.T., Schmidt, G.A., Mann, M.E., Rind, D., and Waple, A., Solar forcing of regional climate change during the Maunder Minimum, Science, 294, 2149-2152, 2001.

Smerdon, J. E., Kaplan, A., Chang, D., and Evans, M. N.,: A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium, J. Clim., 23, 4856–4880, 2010.

Smerdon, J. E.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments, WIRES Climate Change, 3, 63-77 doi: 10.1002/wcc.149, 2012.

Smerdon, J.E, Cook, B.I., Cook, E.R., and Seager, R.: Bridging past and future climate across paleoclimatic reconstructions, observations, and models: a hydroclimate case study, J. Clim., 28(8), 3212-3231, 2015a.

Smerdon, J.E., S. Coats, and T.R. Ault, Model-Dependent Spatial Skill in Pseudoproxy Experiments Testing Climate Field Reconstruction Methods for the Common Era, Climate Dynamics, (submitted), 2015b.

Steiger, N. J., Hakim, G., Steig, E. J., Battisti, D. S. and Roe, G. H.: Assimilation of time-averaged pseudoproxies for climate, J. Clim., 27, 426-441, 2014.

Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene. Geophysical Research Letters, 36(19), doi: 10.1029/2009GL040142, 2009

Stenchikov, G., Hamilton, K., Stouffer, R. J., Robock, A., Ramaswamy, V., Santer, B., and Graf, H. F.: Arctic Oscillation response to volcanic eruptions in the IPCC AR4 climate models, J. Geophys. Res.-Atmos., 111, D07107, doi:10.1029/2005JD006286, 2006.

Stothers, R.B.: The great Tambora eruptions in 1815 and its aftermath. Science, 224(4654), 1191-1198, 1984.

Sundberg, R., Moberg, A., Hind A.: Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 1: Theory, Clim. Past 8, 1339-1353, doi:10.5194/cp-8-1339-2012, 2012.

Taylor, K. E.,. Stouffer, R. J, and Meehl, G. A.: An overview of CMIP5 and the experiment design,, Bull. Amer. Meteor. Soc., 93, doi:10.1175/BAMS-D-11-00094.1., 2012.

Tett, S. F. B., Betts, R., Crowley, T. J., Gregory, J., Johns, T. C., Jones, A., Osborn, T. J., Ostrom, E., Roberts, D. L., and Woodage, M. J.: The impact of natural and anthropogenic forcings on climate and hydrology since 1550, Clim. Dyn., 28(1), 3–34, 2007

Tierney, J. E., Abram, N. J., Anchukaitis, K. J., Evans, M. N., Giry, C., Kilbourne, K. H., Saenger, C. P., Wu, H. C., and Zinke, J.: Tropical sea-surface temperatures for the past four centuries reconstructed from coral archives, Paleoceanography,30, 226–252, doi:10.1002/2014PA002717, 2015.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic reconstructions, Quat. Science Rev., 35, 1-22, 2012.

Tingley, M. P., Stine, A. R., and Huybers, P. : Temperature reconstructions from tree-ring densities overestimate volcanic cooling, Geophys. Res. Lett., 41, 7838–7845, doi:10.1002/2014GL061268, 2014.

Thomson, D.J.: Spectrum estimation and harmonic analysis, Proceed. IEEE,. 70,1055-1096, 1982.

Vieira, L. E. A., Solanki, S. K., Krivova, N. A., and Usoskin, I.: Evolution of the solar irradiance during the Holocene, Astron. Astrophys., 531, A6, doi: 10.1051/0004-6361/201015843, 2011.

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J.E., and  Rajaratnam, B.: Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, Clim. Past, 10, 1-19, doi:10.5194/cp-10-1-2014, 2014.

Wang,J., J. Emile-Geay, D. Guillot, N. P. McKay, and B. Rajaratnam (2015), Fragility of reconstructed temperature patterns over the Common Era: Implications for model evaluation, Geophys. Res. Lett., 42, 7162–7170, doi:10.1002/2015GL065265.

Wang, Y. M., Lean, J. L., and Sheeley Jr, N. R.: Modeling the sun's magnetic field and irradiance since 1713, Astrophys. J., 625(1), 522. Doi: 10.1086/429689, 2005.

Widmann, M., Goosse, H., van der Schrier, G., Schnur, R., and Barkmeijer, J.: Using data assimilation to study extratropical Northern Hemisphere climate over the last millennium. Climate of the Past 6, 627–644, 2010

Wilmes, S., Raible, C. and Stocker, T.: Climate variability of the mid- and high-latitudes of the Southern Hemisphere in ensemble simulations from 1500 to 2000AD, Clim. Past 8: 373–390, 2012.

Wigley, T. M. L., Ammann, C. M., Santer, B. D., and Raper, S. C. B.: Effect of climate sensitivity on the response to volcanic forcing, J. Geophys. Res., 110, D09107, doi:10.1029/2004JD005557, 2005.

Wunsch, C.: The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations, Bull. Amer. Meteor. Soc., 80, 245-255, 1999.

Yoshimori, M., Stocker, T. F., Raible, C. C., and Renold, M.: Externally-forced and internal variability in ensemble climate simulations of the Maunder Minimum, J. Climate, 18, 4253-4270, 2005.

Zanchettin,D., Bothe, O., Lehner, F., Ortega, P., Raible, C. C., and Swingedouw, D.: Reconciling reconstructed and simulated features of the winter Pacific-North-American pattern in the early 19th century, Clim. Past, submitted, 2015.

Zorita, E., Gonzalez-Rouco, F., and Legutke, S.: Testing the Mann et al. (1998) approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model, J. Climate, 16, 1378-1390, 2003.

Zunz V., Goosse, H. and Massonnet, F.: How does internal variability influence the ability of CMIP5 models to reproduce the recent trend in Southern Ocean sea ice extent? The Cryosphere 7, 451–468, 2013.

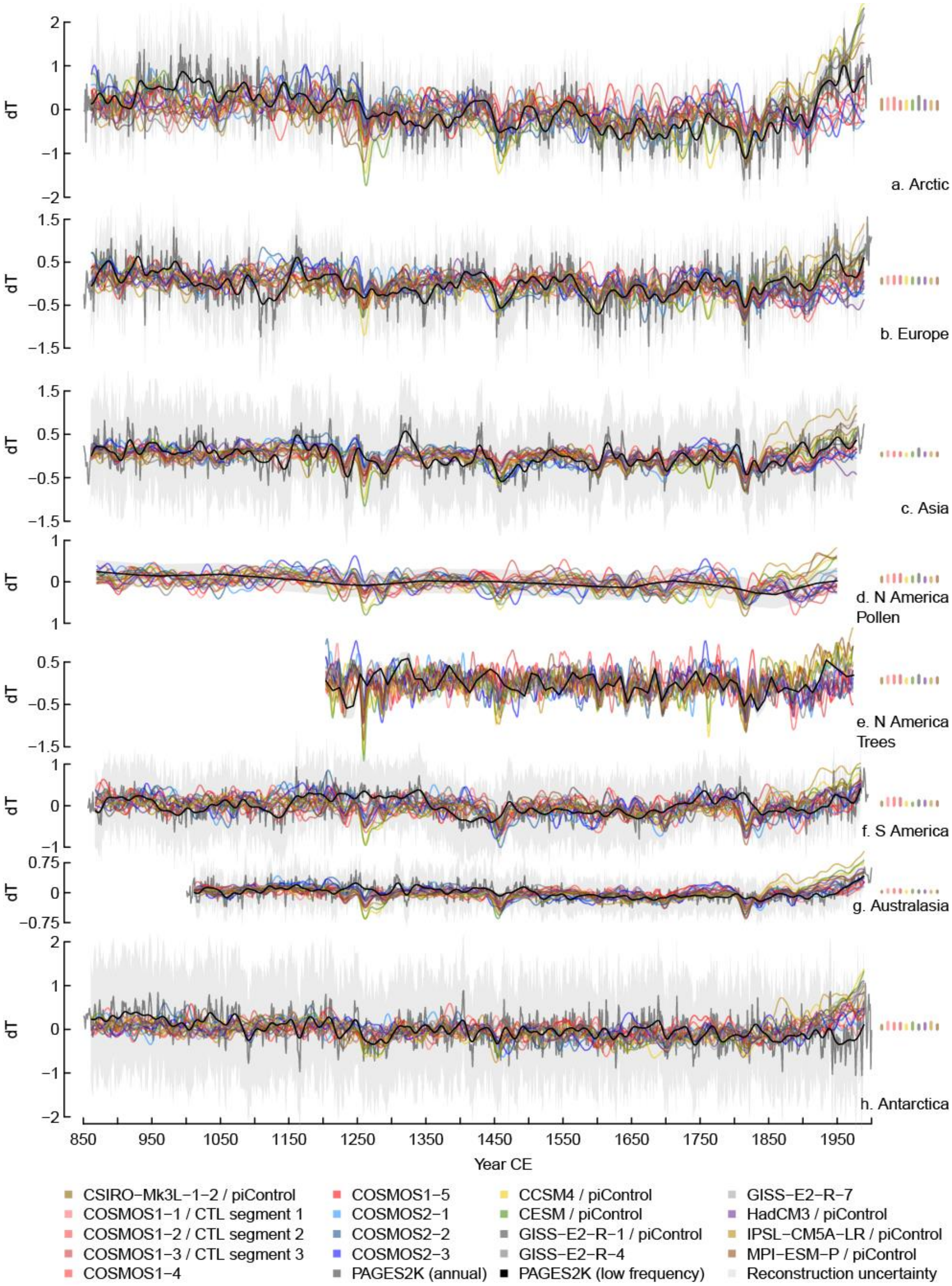


Figure 1: Series of simulated temperatures and PAGES 2k reconstructions for the seven continent-scale regions. The reconstructions are shown at their original resolution and after a smoothing using a 23-year Hamming filter, except for the North American reconstructions. Only the smoothed series are shown for models. Grey shading denotes each reconstruction's original uncertainty estimates. Segments on the right indicate the unforced variability of the 23-year Hamming filtered times series in the respective control simulations (standard deviation of the time series, colours as in the caption). The anomalies are computed compared to the mean of the time series over the full length of temporal overlap between simulations

and reconstruction. Note the different scales in the y-axis of the various regions.

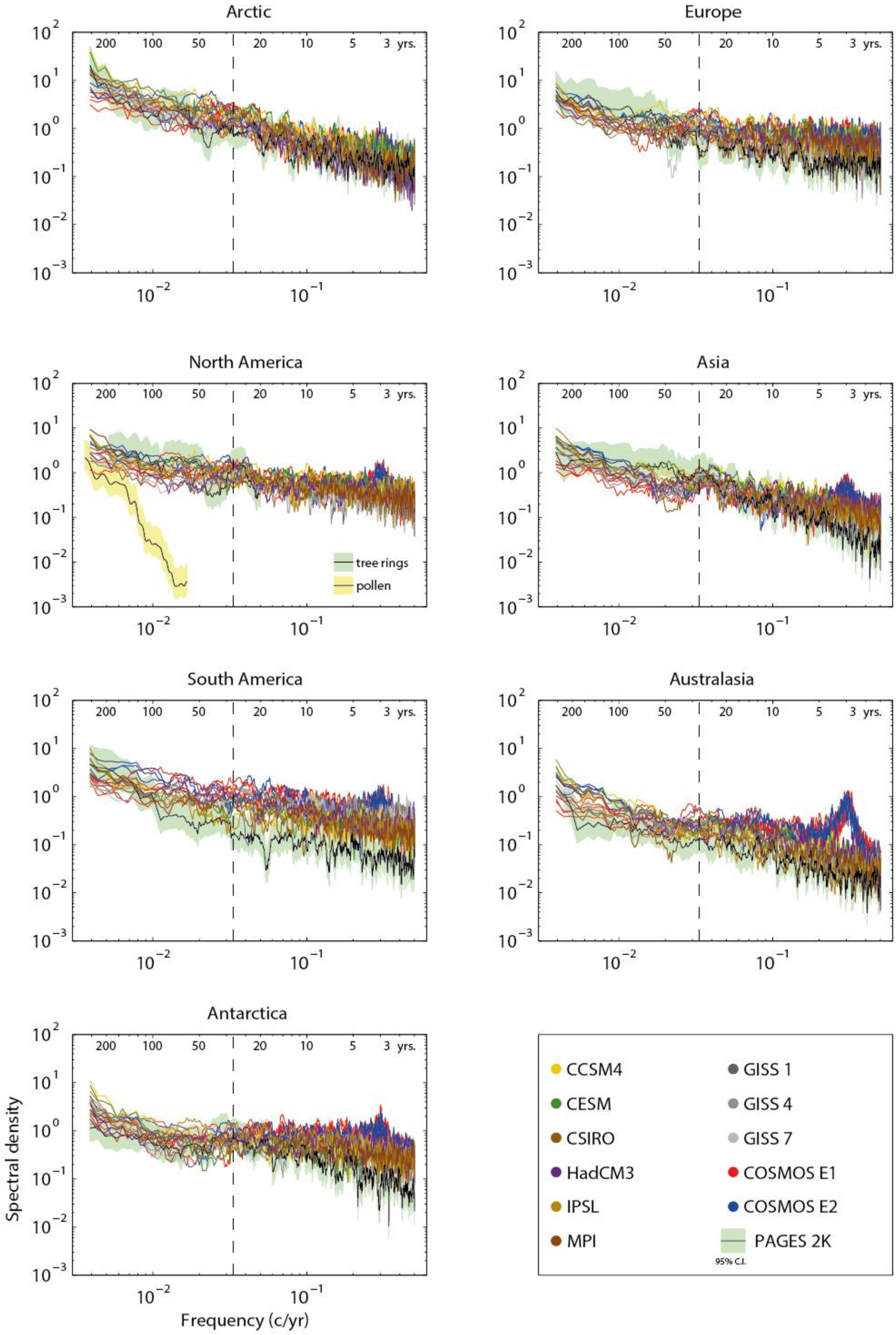
Arctic
Europe
North America
Asia
South America
Australasia
Antarctica
200 100 50 20 10 5 3 yrs.
tree rings
pollen
Spectral density
Frequency (c/yr)
CCSM4
CESM
CSIRO
HadCM3
IPSL
MPI
GISS 1
GISS 4
GISS 7
COSMOS E1
COSMOS E2
PAGES 2K
95% C.I.

Figure 2: Spectral densities for simulations and reconstructions for PAGES 2k regions, calculated using all existing data in the period 850-2005 CE. Reconstruction spectra are illustrated with their 95% confidence intervals in coloured bands, while model spectra are shown with single coloured lines. Dashed vertical lines denote the limit for frequencies and periods of relevance (to the left of the line) for analyses made at the 15-year resolution, or with a 23-point Hamming window, as in many other analyses in this study. The multi-taper method (Thomson, 1982; Percival and Walden, 1993) was used, with the time-bandwidth product set to 4 and with long-term averages subtracted before estimating the spectra. Units are temperature variance (°$C^2$ or $K^2$) per frequency (c/year).

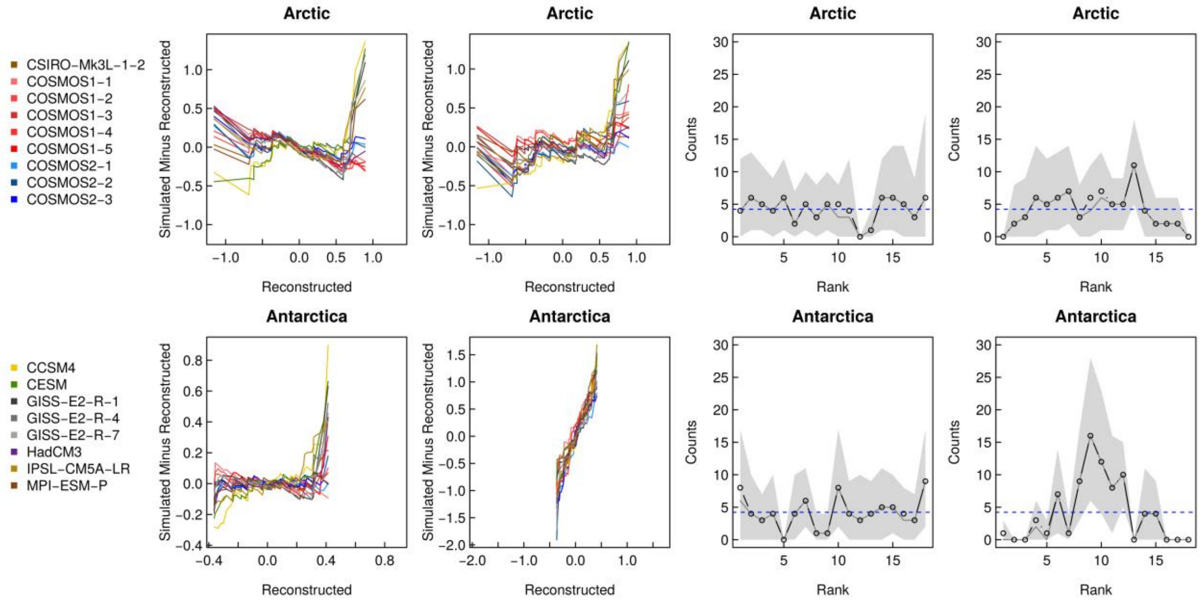


Figure 3: Climatological consistency (first two columns): residual quantile-quantile plots for the full period; and probabilistic consistency (last 2 columns): rank counts for the full period. The top row is for the Arctic, and the bottom row is for Antarctica. For both the climatological and probabilistic consistency, the computations are obtained by neglecting the uncertainties (left plot) and using the uncertainties provided with the original reconstructions (right plot). For the climatological assessment, positive and negative slopes or large differences from 0 emphasize lack of consistency. For the probabilistic measure, U- or dome-shaped features highlight lack of consistency.

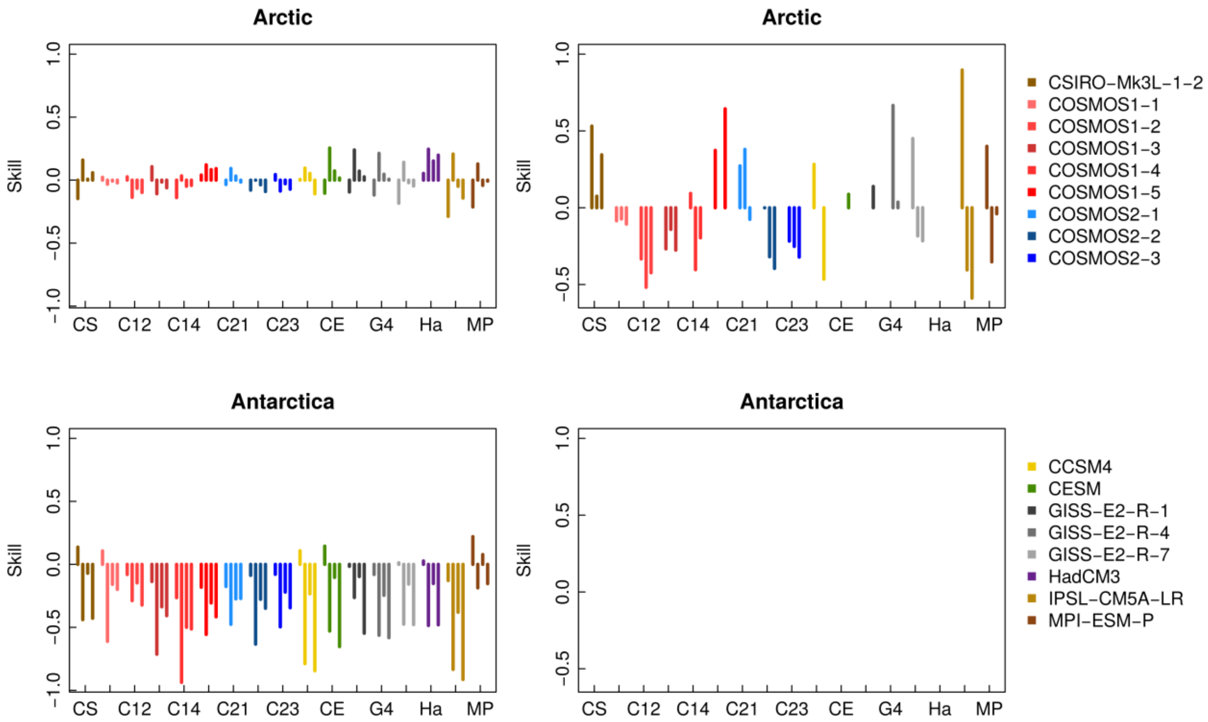


Figure 4: Skill metric for the individual models for all periods (from left to right: 850-1350, 1350-1850, 850-1850, 850-2000). Top row for the Arctic, bottom for Antarctica. The computations assume no uncertainties (left plot) and uncertainties provided with the original reconstructions (right plot). When the skill is undefined (as for Antarctica when using the original error estimates) no bar is shown. Positive values indicate skill in this simple evaluation.

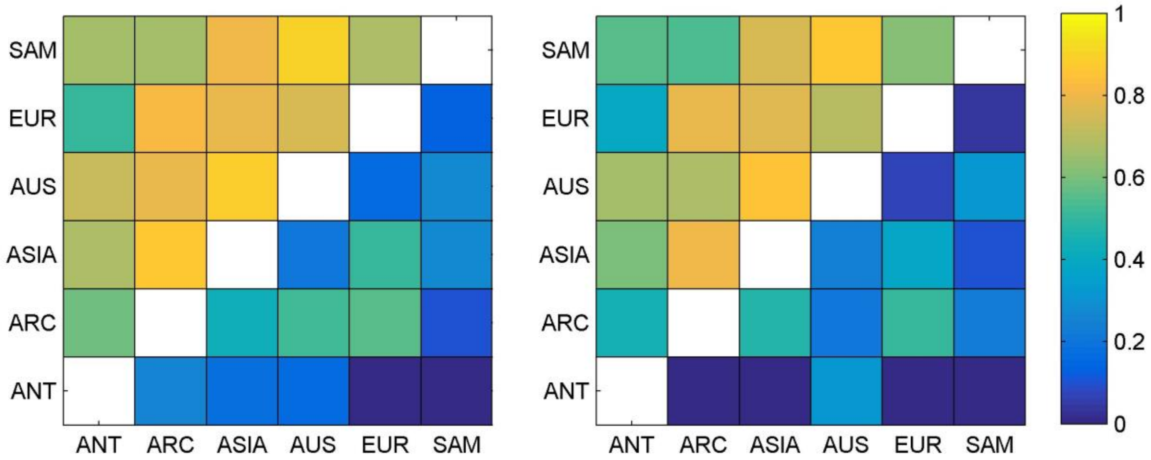


Figure. 5: Correlations among the PAGES2k regions for detrended simulated and reconstructed time series filtered using a 23-year Hamming filter. Left-hand panel: forced simulation with MPI-ESM (upper triangle) PAGES 2k reconstructions (lower triangle) for 1012-1978 CE. Right-hand panel: forced simulation with MPI-ESM for the preindustrial period 1012 CE – 1850 CE (upper triangle) and unforced control simulation with MPI-ESM (lower triangle).

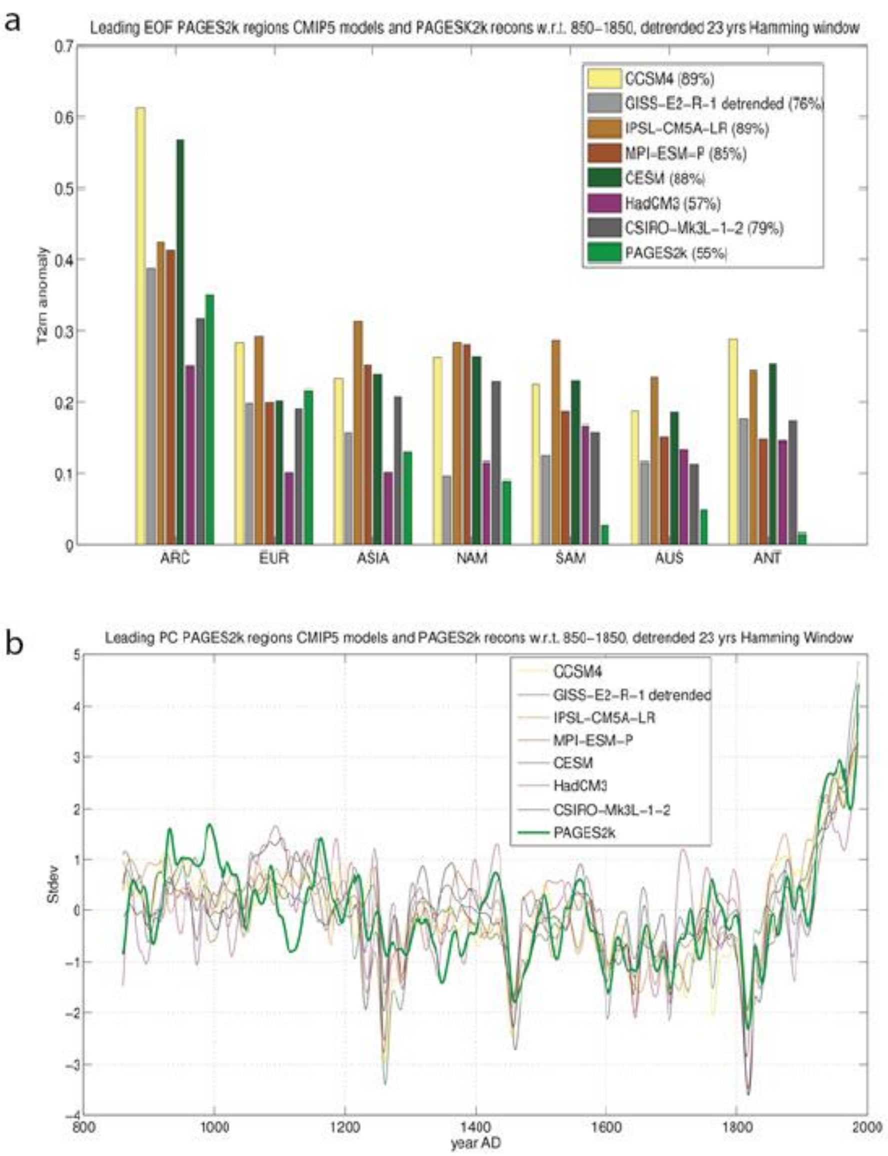


Figure 6: a) Leading EOFs of the near-surface temperature simulated by each CMIP5/PMIP3 model and reconstructions over the full period 850–2004 CE. The EOF analysis is based on the covariance matrix with respect to temperature anomalies for the pre-industrial period 850–1850 CE. Values in parentheses correspond to the amount of variance represented by the leading EOF. b) Time series of the principal components (PCs) corresponding to the leading EOF for the PMIP3 simulations and PAGES2k reconstructions. The time series were filtered with a 23-year Hamming filter and were linearly detrended before the covariance matrix was calculated. The PC time series are shown as standardized anomalies from the average over the full period 850–2004 CE. Positive PC values correspond to positive temperature anomalies in the respective regions. Results for single member realizations and the pre-industrial period are presented in the Figures S7 and S8, respectively.

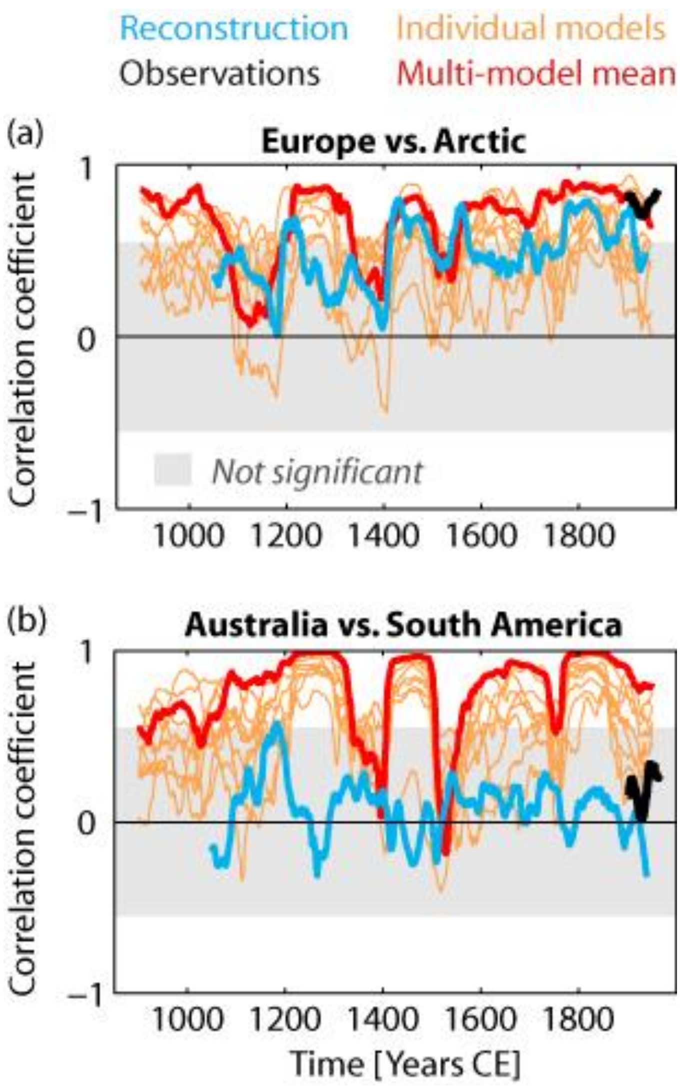


Figure 7: 100-year moving Tukey window correlations between selected PAGES 2k regions for the PAGES 2k reconstructions (blue) and PMIP3 models (8 models in orange, multi model mean in red) and observations from HadCRUT4 (Morice et al., 2012, black). Each 100-year segment is linearly detrended beforehand. Grey shading illustrates correlations that are not significant at the 5% level. (a) Correlation between Arctic and Europe as an example of good agreement of model and reconstruction, (b) correlation between Australia and South America as an example of poor agreement. For all other combinations see Figure S9.

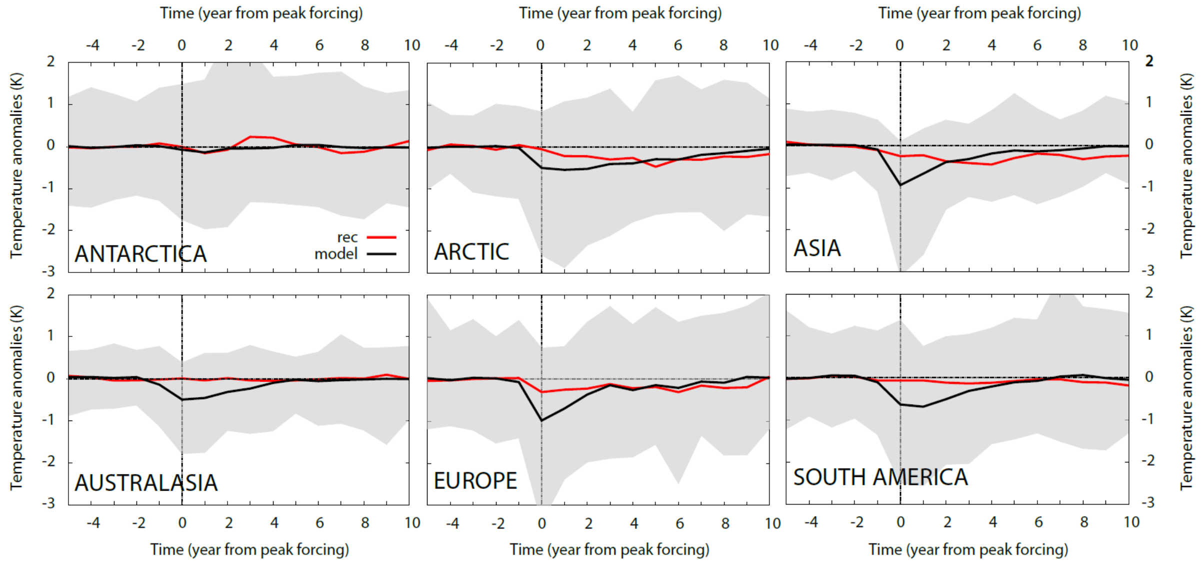


Figure 8. Superposed Epoch Analysis of the impact of the volcanic activity in the reconstructed and simulated temperatures. Superposed composites of temperature responses during selected periods when peak negative forcing in the Crowley and Unterman (2012) volcanic reconstruction are aligned. The composite is produced by selecting the 12 strongest volcanic events, starting 5 years before the date of the peak eruption and ending 10 years after the event. Each panel indicates the reconstructed (red lines) and simulated (black) composites of the temperature response for each Pages2k region. The grey shading indicates the complete range of simulated temperature responses.

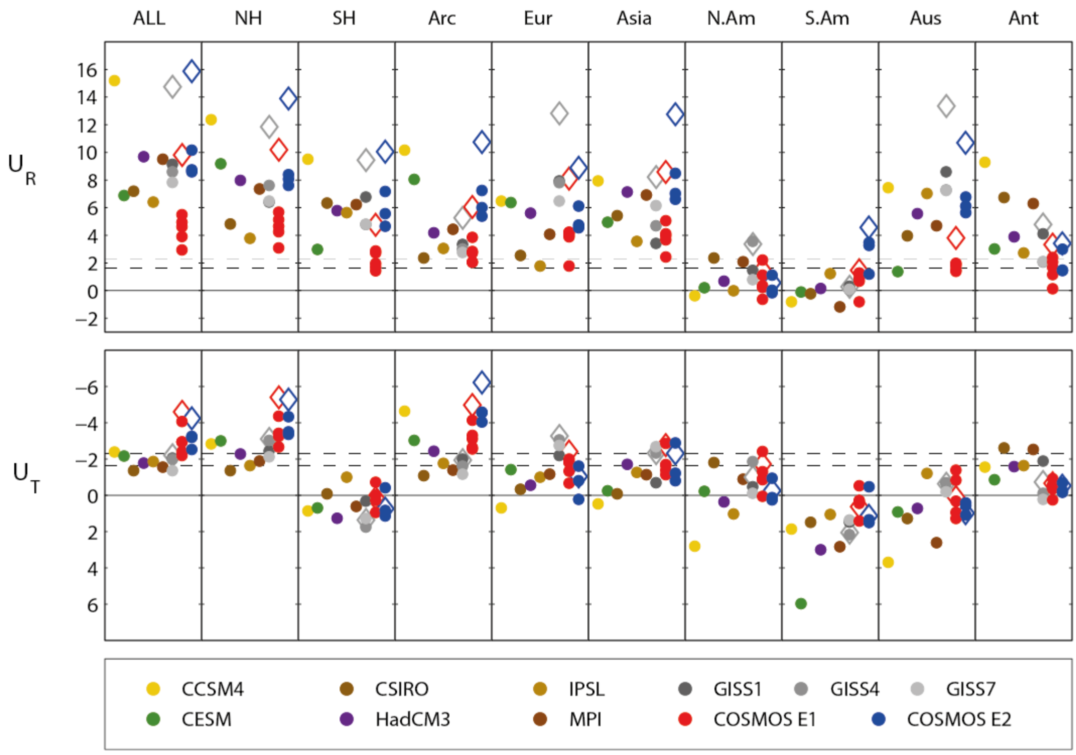


Figure 9: Correlation ($U_R$) and distance ($U_T$) statistics for PAGES 2k regions, with hemispheric and global combinations of all regional data, in the period 861–1850 CE. Positive $U_R$ indicates that simulations and reconstructions have a positive correlation and that they share an effect of temporal changes in external forcings. Negative $U_T$ indicates that a forced simulation is closer to the observed temperature variations than its own control simulation. The analysis reveals a notably better general agreement between simulations and reconstructions for the Northern Hemisphere as compared to the Southern Hemisphere. 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.

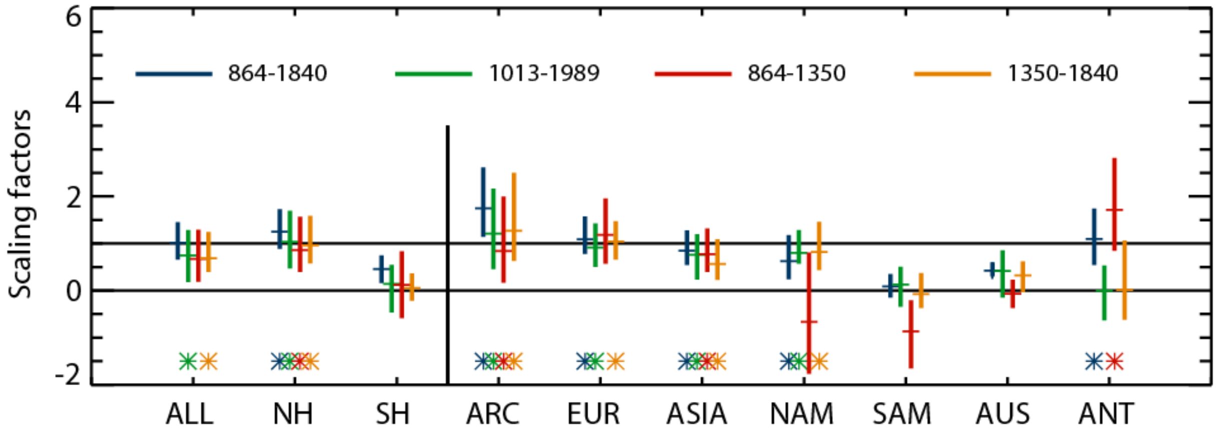


Figure 10: Detection and attribution results for PAGES 2k regions. Vertical bars indicate 5-95% scaling factor ranges, with a cross indicating the best fit. Scaling factors that are significantly offset from ‘0’ indicate that the response to forcing is detected, and those that encompass ‘1’ indicate that the magnitude of the forced response agrees with simulations. For each region scaling ranges are shown for four different time periods (colours). For the Northern hemisphere (NH), Southern Hemisphere (SH) and global (ALL), the regressions were carried out on the combined data from all the applicable regions. An asterisk indicates that the detection analysis has been successful, namely the forced response is significantly greater than zero and that the residuals are consistent with model-based samples of internal variability.

Table 1. Description of the model simulations

| Model | # runs | Resolution | Resolution | Forcing | | | | | | Reference | piControl length (yrs) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | S | V | G | A | L | O | | |
| CCSM4 | 1 | 0.9° x 1.25°, L26 (atm) nominal 1°, L60 (ocn) | 288 x 192, L26 (atm) 320 x 384, L60 (ocn) | 10 | 20 | 30, 31, 32 | 40 | 50 | 60 | Landrum et al. (2013) | 500 |
| CESM1 | 1 | 0.9° x 1.25°, L26 (atm) nominal 1°, L60 (ocn) | 288 x 192, L26 (atm) 320 x 384, L60 (ocn) | 11 | 20 | 30, 31, 32 | 40 | 50 | 1990 CE | Lehner et al. (submitted.) | 465 |
| CSIRO-Mk3L-1-2 | 1 | 5.63° x 3.21°, L18 (atm) 2.81° x 1.61°, L21 (ocn) | 64 x 56, L18 (atm) 128 x 112, L21 (ocn) | 12 | 21 | 30, 31, 32 | none | none | 60 | Phipps et al. (2013) | 1150 |
| GISS-E2-R | 3 | 2° x 2.5°, L40 (atm) 1° x 1.25°, L32 (ocn) | 144 x 90, L40 (atm) 288 x 180, L32 (ocn) | 12 | 21, 20 | 30, 31, 32 | 40 | 50, 51 | 60 | Schmidt et al. (2014b) | 1162 |
| HadCM3 | 1 | 3.75° x 2.46°, L19 (atm) 1.25° x 1.25°, L20 (ocn) | 96 x 73, L19 (atm) 288 x 144, L20 (ocn) | 12 | 21 | 30, 33, 32 | 41 | 51 | 60 | Schurer et al. (2013) | 1199 |
| IPSL-CM5A-LR | 1 | 3.75° x 1.88°, L17 (atm) 1.98° x 1.21°, L32 (ocn) | 96 x 96, L17 (atm) 182 x 149, L32 (ocn) | 10 | 22 | 30, 31, 32 | none | none | 60 | Dufresne et al. (2013) | 1004 |
| MPI-ESM-P | 1 | 1.84° x 1.84°, L47 (atm) nominal 1.5°, L40 (ocn) | 196 x 98, L47 (atm) 256 x 220, L40 (ocn) | 10 | 21 | 30, 31, 32 | 40 | 52 | 60 | Jungclaus et al. (2014) | 1155 |
| ECHAM5/MPIOM (COSMOS) | E1:5 | 3.75° x 3.75°, L19 (atm) nominal 3°, L40 (ocn) | 96 x 48, L19 (atm) 120 x 101, L40 (ocn) | 13 | 21 | 32, 34 | 40 | 52 | 61 | Jungclaus et al. (2010) | 1000 |
| | E2: 3 | | | 14 | 21 | 32, 34 | 40 | 52 | 61 | | |

**Forcings**: S,V,G,A, L and O stands respectively for Solar, Volcanic, Greenhouse gas, Aerosols, Land use and Orbital forcing, respectively, derived from the following references:

10 = Vieira and Solanki (2010) spliced to Wang et al. (2005)
11 = as 10, but scaled to double the Maunder Minimum-Present Day amplitude
12 = Steinhilber et al. (2009) spliced to Wang et al. (2005)
13 = Krivova et al. (2007)
14 = Bard et al. (2000)
20 = Gao et al. (2008)
21 = Crowley and Unterman (2013)
22 = Ammann et al. (2007)
30 = Flückiger et al. (1999, 2002); Machida et al. (1995)
31 = Hansen and Sato (2004)
32 = MacFarling Meure et al. (2006)
33 = Johns et al. (2003)
34 = $CO_2$ diagnosed by the model.
40 = Lamarque et al. (2010)
41 = Johns et al. (2003)
50 = Pongratz et al. (2009) spliced to Hurtt et al. (2011)
51 = Kaplan et al. (2011)
52 = Pongratz et al. (2008)
60 = Berger (1978)
61 = Bretagnon and Francou (1988)