Clim. Past Discuss., 11, 2483–2555, 2015
www.clim-past-discuss.net/11/2483/2015/
doi:10.5194/cpd-11-2483-2015
© Author(s) 2015. CC Attribution 3.0 License.

Climate of the Past Discussions
Open Access

This discussion paper is/has been under review for the journal Climate of the Past (CP).
Please refer to the corresponding final paper in CP if available.

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

PAGES2k-PMIP3 group*

* A full list of authors and their affiliations appears at the end of the paper.

Received: 06 May 2015 – Accepted: 01 June 2015 – Published: 29 June 2015

Correspondence to: H. Goosse (hugues.goosse@uclouvain.be)

Published by Copernicus Publications on behalf of the European Geosciences Union.

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 the collection of state-of-the-art model simulations driven by realistic external forcings following the PMIP3 protocol are jointly analysed. The first aim is to estimate the consistency between model results and reconstructions for each continental-scale region over 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 latitudes 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.

CPD
11, 2483–2555, 2015

1 Introduction

The past millennium is an important period for testing our understanding of the mechanisms that give rise to the climate system variability (e.g., Masson-Delmotte et al., 2013). Constraints and uncertainties on 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 places 20th-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, 2000; Cook et al., 1999, 2004, 2010; 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).

A comprehensive comparison of reconstructions and realistically forced model simulations in paleoclimatology has several important outcomes. Reconstruction of past events provides out-of-sample tests of the fidelity of modelled processes and their role in explaining past climatic variations. Reconstructions and simulations can also be used jointly to validate 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 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., 2013a, b; 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

CPD
11, 2483–2555, 2015

forced signal is extracted from realistically forced experiments (Schurer et al., 2013). Multivariate spatiotemporal comparisons of observations and reconstructions with realistically forced simulations have the potential to identify processes within simulations that are most consistent with observations.

Separately, simulations can provide the basis for observing network array design (Comboul et al., 2015), producing hypotheses testable with future observations. Simulation results 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 univocal 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 (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). The uncertainty in paleoclimatic reconstructions is also not always well understood 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

CPD
11, 2483–2555, 2015

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

CPD
11, 2483–2555, 2015

also simply reflect a different sample of internal variability (defined here as the fraction of climatic variability that cannot be explained by changes in external forcings). Indeed, comprehensive climate models have their own internal climate variability that depends on the processes incorporated, the numerical schemes implemented in the code and the initial conditions applied in a specific experiment. If a model represents the real world in a satisfactory way, the observed trajectory is just 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) or a set of models (e.g., Jansen et al., 2007; Lehner et al., 2012; Fernandez-Donado et al., 2013; Bothe et al., 2013b; Moberg et al., 2015). 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.

Recently, 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). These two periods are often referred to as the Medieval Climate Anomaly (MCA), and the Little Ice Age (LIA), respectively, although their exact timing has been

CPD
11, 2483–2555, 2015

debated and the adequacy of their names has been questioned (e.g., Jones and Mann, 2004). 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; 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 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, has also been estimated using detection and attribution techniques. The results of these studies show that the response to volcanic eruptions can be clearly detected in reconstructions, but 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

CPD
11, 2483–2555, 2015

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 there is also weak evidence for a contribution from a small 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., 2013) and regional temperature variability (Luterbacher et al., 2004; Hegerl et al., 2011; Goosse et al., 2012a; Gergis et al., 2015). Additionally, the long-term behaviour of Southern Hemisphere circulation features such as the Southern Annular Mode (Wilmes et al., 2012; Abram et al., 2014) and the El Niño–Southern Oscillation (ENSO, Mann et al., 2009; Ault et al., 2013; Borlace et al., 2013; Landrum et al., 2013; Tierney et al., 2015) were investigated. 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, noting that an extended cold period (1594–1677 CE) was observed in both hemispheres, but no globally coherent warm phase occurred during the pre-industrial (1000–1850 CE) period. 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). Comparison with the 24-member climate model ensemble suggested that the simulations underestimate the influence of internal variability, particularly in the ocean-dominated Southern Hemi-

CPD
11, 2483–2555, 2015

sphere (Neukom et al., 2014) while Schurer et al. (2013) found the residual Northern Hemisphere mean temperature variability after subtracting forced responses to be broadly consistent with variability in model control simulations, particularly for the better reconstructed period following the MCA.

While several studies provide 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 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 external forcings selected in the PMIP3 framework (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 presents a brief overview of the PAGES 2k reconstructions and PMIP3 simulations analysed here. We use several statistical methods to answer each of the key science questions to achieve robust results. Their specific implementation is described in Sect. 3. Next, each continental-scale region is studied separately to determine whether the reconstructed and simulated time series have similar characteristics,

CPD
11, 2483–2555, 2015

in terms of the magnitude and timing of the observed changes as well as the spectral distribution of the variance (Sect. 4). Section 5 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 then presented in Sect. 6. Section 7 provides a discussion of our results, their limitations and how our conclusions compare to previous studies. Finally, Sect. 8 summarizes the main findings and provides perspectives for future developments. Several additional analyses are provided as Supplement for completeness and further reference.

2 Data

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

CPD
11, 2483–2555, 2015

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 is 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 region also used individually selected approaches to assess the uncertainty of their temperature reconstructions. They are generally assumed to be constant in time while in reality they are likely to be non-stationary, with different high- and low-frequency variations.

2.2 Climate model simulations

The climate model simulations used in this study are listed in Table 1, summarizing model specifications such as resolution, forcings 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 from a suite of available pre-PMIP3 and PMIP3 simulations that fulfilled the following criteria: (i) they run continuously from 850–2000 CE;

CPD
11, 2483–2555, 2015

(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. Most simulations comply with 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 %. Additionally, 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 variable extracted from the simulation outputs is the monthly mean surface air temperature (labelled "tas" in the Climate Model Output Rewriter framework). 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).

3 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 reconstructions can never be expected.

CPD
11, 2483–2555, 2015

The first group of methods used in this study assess if temperature reconstructions have similar statistical properties as simulations, including similar variance, without specifying the forcing role (e.g., Sects. 3.1 and 3.2), while others focus on detecting and quantifying the contribution of forcing (Sects. 3.3–3.6). Several techniques are relatively standard while the specific implementation of some of the methods will be introduced below, prior to their application.

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 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, or other technical restrictions.

3.1 Probabilistic and climatological consistency

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. The two concepts of probabilistic and climatological consistency can be seen as two alternative ways to evaluate biases and spreads in simulations and reconstructions. In the current study, however, the analyses use temperature anomalies from long-term averages, and

CPD
11, 2483–2555, 2015

hence the bias is always zero by construction. Thus, our analysis mainly assesses two different aspects of spread in the distributions of the regional temperature reconstructions and climate model simulations.

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

Probabilistic consistency refers to the position of the reconstruction in the range spanned by the ensemble of simulations. This position can be first calculated for every time interval of chosen length, to later derive aggregated measures of consistency over time. To assess probabilistic consistency, we test whether the occurrence frequencies of the simulation ensemble agree with those of the verification target, within limits of uncertainty. At each time step, we identify the rank of the temperature reconstructions within the set formed by the combination of the simulation ensemble and those temper-

CPD
11, 2483–2555, 2015

ature reconstructions (Anderson, 1996). Histograms of the ranks should be flat under exchangeability, 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). Flatness of the histograms is thus a necessary condition for our simulation ensemble to be considered as a reliable representation of the target. The histograms visually highlight biases (meaning here an offset in mean between the target and the ensemble) and differences in ensemble variance. Over-dispersion (ensembles that are too wide) and under-dispersion (ensembles that are too narrow) are identified by dome- or U-shaped histograms, respectively. Such shapes imply that the target data are too often close to the central rank or too often on the outer ranks (i.e., far from the mean of the ensemble of simulations). Slopes in the histograms reveal biases, with positive (negative) slopes suggesting the target data are ranked high (low) too often.

We assess probabilistic and climatological consistency 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.

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.

3.2 Skill

In Sect. 4.4, 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

CPD
11, 2483–2555, 2015

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 in Sect. 4.4) 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.

3.3 Superposed epoch analysis

In Sect. 6.1, the response of the PAGES 2k reconstructions and the various model simulations to external forcing from solar and volcanic activity is evaluated using a superposed epoch analysis approach. This is done by generating composites of reconstructed and simulated temperature sequences corresponding to the timing of the strongest volcanic events. 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 response to volcanic aerosol forcing is evaluated at interannual and multidecadal time scales for two different external forcing estimates (Gao et al., 2008; Crowley and

CPD
11, 2483–2555, 2015

Untermann, 2013) that have been used as last-millennium boundary conditions in the PMIP3-CMIP5 simulations (Schmidt et al., 2011, 2012).

The volcanic composite at interannual timescales is generated by first selecting the 12 strongest volcanic events. The mean from 5 years before to 10 years after the date of the peak eruption is then computed for the forcing sequence as well as for the simulated and reconstructed temperature sequences. In the case of the multidecadal composites, the time series are first filtered with a 40 year low pass filter using least-squares coefficients (Bloomfield, 1976). For the composite, the 5 strongest events are selected and the means from 40 years before to 40 years after the eruption are calculated. All the events are individually selected for each of the PAGES 2k regions making use of the latitudinal discretization of the volcanic forcing (see Supplement Sect. S2).

The composites for the strongest multidecadal changes in the solar forcing are based on 80 year time windows centred on the seven years with largest solar forcing of the reconstruction of Amman et al. (2007), which was band-pass filtered to extract variability at periods between 20–160 years prior to analysis (Masson-Delmotte et al., 2013)

3.4 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, henceforth SUN12), Hind et al. (2012) and Moberg et al. (2015). The statistical model underlying this framework has similar components to the one used in detection and attribution studies (see Sect. 3.5), but there are some differences. An important similarity is the idea that temperature variations can be expressed as a sum of forced and unforced variability. The two frameworks explicitly distinguish internal variability in simulations and in observations, which can consist of instrumental observations or, as in this study, proxy-based climate reconstructions. The SUN12 framework also explicitly accounts for error variance in the observations, such as non-climatic noise in proxy data. It even allows this type of error to vary with time, if such information is available. Despite simi-

CPD
11, 2483–2555, 2015

larities in the underlying assumptions, the main purposes of the SUN12 and detection and attribution approaches differ. While detection and attribution studies seek to identify the forced response in observations, the SUN12 framework was developed as tool for evaluating forced simulations, with the aim of testing if one simulation significantly fits observations better than another simulation or to rank a set of plausible simulations. In the current study, this framework is mainly used to investigate the common behaviour of all simulations by means of how well they agree with the different regional reconstructions.

The SUN12 framework 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.

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

CPD
11, 2483–2555, 2015

low solar forcing, and for the GISS ensemble, metrics are calculated for each individual simulation and for the entire ensembles, following Moberg et al. (2015).

All records were recalibrated against their instrumental target temperature time series (see Supplement Sect. S1) 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). Note that this recalibration is specific for the calculation of U_R and U_T statistics and is not applied for any other diagnostics. For simplicity, it is assumed here that each proxy record has the same statistical precision in its entire length, despite the fact that their precision typically decreases back in time as the number of contributing local proxy series decreases. Therefore, the derived measures are only approximate values, but a more accurate treatment would require detailed work far beyond the scope of this study. As for the methods outlined in Sects. 3.1–3.3, this analysis also uses anomalies from long-term averages to avoid systematic climatological bias influencing the results.

3.5 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) or if external forcing is necessary to explain them. The result is an estimate of the magnitude of the forced response in the reconstruction, with an uncertainty estimate. Detection and attribution is also used to determine the relative contribution by different forcings simultaneously to a period or climatic event, with uncertainty estimates reflecting if the contribution of different forcings can be separated from each other and from climate variability (see Bindoff et al., 2013; Hegerl and Zwiers, 2011). To estimate the different contributions from several individual forcings, it is necessary to have access to separately forced simulations with each individual forcing. This is not the case in this study because all

CPD
11, 2483–2555, 2015

forced simulations were made combining all forcings together; hence we focus here on estimating the magnitude of the overall forced response.

Detection and attribution studies often rely on multiple regression of reconstructions onto the response expected by different individual contributing forcings, assuming that climate models approximately capture the response to individual forcings in shape (e.g., pattern in time or spatial pattern of the response), but may misrepresent the magnitude of the overall response. This is a reasonable assumption since the magnitude of the response to forcings is affected by uncertainty in the transient climate sensitivity. Moreover, the magnitude of forcings itself is also often uncertain, such as for the low-frequency component of solar forcing (see e.g., Schmidt et al., 2011, 2012). In contrast, spatial patterns of response are driven by the dynamics and thermodynamics of the climate system, such as the different thermal properties for the land–sea contrast in temperature changes. The temporal pattern of response that is used in this study is driven by the forced temporal pattern, modified by the delay in response caused by the thermal inertia of the climate system. While aspects of the temporal pattern are uncertain, such as the relative magnitude of volcanic eruptions relative to each other, the timing of eruptions and magnitude of overall volcanic forcing is more important for identifying any forced signal (see Hegerl et al., 2006). A difficulty in the application of detection and attribution methods to reconstructions is accounting for uncertainty in both reconstructions and forcings. This can be addressed to some extent by using multiple reconstructions and forcing estimates (e.g., Schurer et al., 2014), but a more systematic approach is desirable.

The detection and attribution framework applied here has been extensively used for instrumental data (Bindoff et al., 2013) and to some extent for paleoclimatic reconstructions (see Hegerl et al., 2007; Schurer et al., 2014). This approach calculates a possible scaling range for the response to the external forcing in the reconstruction

CPD
11, 2483–2555, 2015

(Eq. 2) based on total least squares regression (Allen and Stott, 2003):

$$Y(t) = \sum_{i=1}^{m} \beta_i (X_i(t) - \gamma_i(t)) + \gamma_0(t) \quad (2)$$

where Y, the reconstructed temperature, is equal to a linear combination of m different model fingerprints X_i (where m in this analysis is always equal to 1 as only the response to all the forcings together is analysed here) multiplied by a scaling factor β_i. Each model simulation has associated internal variability γ_i and the reconstructions contain a realization of internal variability γ_0. The scaling factors β_i determine the amplitude of the fingerprints in the reconstructions. A range of scaling factors is obtained using samples of internal variability taken from model simulations. A forcing is said to be detected if a scaling value of zero is rejected at some significance level, for example, the 5 % level. To evaluate the self-consistency of the regression results, the residual of the fit is checked against estimates of model-based internal variability. This is the same method as used in Schurer et al. (2013).

4 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 Supplement (Fig. S1 and Sect. 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 are compared and discussed in more detail in Sect. 7.

CPD
11, 2483–2555, 2015

4.1 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 estimating the standard deviation of control runs in each model. The unforced variability is generally similar in all models in all the regions, with weaker amplitudes in Australasia and Asia.

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 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 19th century eruptions are clearly reflected in many regions, but 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 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 Supplement 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

CPD
11, 2483–2555, 2015

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 20th century. Although the European temperature reconstruction and simulated series disagree substantially in some parts of the 12th 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.

CPD
11, 2483–2555, 2015

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 20th century while the simulations show nearly no long-term cooling but agree on the warming onset in the beginning of the 20th century.

4.2 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

CPD
11, 2483–2555, 2015

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. Even more pronounced differences of high-frequency variance are 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.

4.3 Consistency estimate

Figures 3, S3 and S4 show the results for the probabilistic and climatological consistency assessment for all regions. The regions selected for Fig. 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 Supplement 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 Sect. 4.1). 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 (Figs. 3 and S3). The simulated quantiles can deviate strongly from the recon-

CPD
11, 2483–2555, 2015

structed 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 Fig. 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 (Figs. 3 and 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.

CPD
11, 2483–2555, 2015

4.4 Skill estimate

Figure 4 presents a skill measure for the Arctic and Antarctica, as an example, with the other PAGES 2k regions displayed in Fig. 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 simulation 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 Sect. 4.3, 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 Sect. 3.2). Additionally, the skill is computed assuming a reduction by a factor $\sqrt{15}$ in the Supplement 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.

5 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 inves-

CPD
11, 2483–2555, 2015

tigate 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.

5.1 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 (Figs. 5 and 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 1012–1978 CE (1000–1990 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 Fig. 5 (see Fig. S6 for the other models). Correlations in the forced MPI-ESM-P simulation for the whole period are very high. 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 Sect. 6 but we must already recall here that, in contrast to other analyses presented above, the evaluation of the spatial correlation do 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 very high correlations for the entire period are likely to be a consequence of the rather homogeneous and

CPD
11, 2483–2555, 2015

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 contribute 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 vs. unforced variability, and in particular to discrepancies in the magnitude of the response to external forcing in the selected models.

5.2 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 Supplement Sect. 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 just a 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 19th and 20th century. Additionally, the influence of volcanic eruptions on reconstructed temperatures is visible during some periods, especially during

CPD
11, 2483–2555, 2015

the mid-13th century (although not in the reconstructions), the mid-15th century and the beginning of the 19th 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 Sect. 5.1, 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 1200 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.

5.3 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, (Figs. 7 and 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 11th and 12th century and in shorter periods around 1500 and 1750. High coherence

CPD
11, 2483–2555, 2015

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 Sect. 5.1). 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 Supplement Sect. S5 and Fig. S11), or that there is more non-temperature noise in the proxy data from those regions.

6 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.

6.1 Superposed epoch analysis

The annual forcing composites, based on the major volcanic events in the Crowley and Unterman (2012) reconstruction are generally similar for six PAGES 2k regions (North America is not analysed here), aside from weaker South American results (Fig. 8). The peak values of the composites are approximately $-4\,W\,m^{-2}$, while the forcing during the largest single events reaches $-10\,W\,m^{-2}$. Despite very similar regional volcanic

CPD
11, 2483–2555, 2015

forcing, the regional temperature response varies considerably in the simulations and in the reconstructions and the composite temperature response is larger in the simulations than in the reconstructions. The largest temperature responses in simulated and reconstructed temperatures are found in Europe and Asia. The composite averages are, however, larger in model results with values of up to −1 °C compared to about −0.25 °C in reconstructions. The Arctic and South America show smaller simulated temperature changes (around −0.5 °C) and the average responses in the reconstructions are even smaller but stay at levels of −0.1 to −0.2 °C during several years. 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. 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 Supplement Sect. S6 and Fig. S11).

At multidecadal timescales, the volcanic forcing lies in the range of −0.5 to −1 $W m^{-2}$ for the Crowley and Unterman (2012) reconstruction and −0.5 to −0.75 $W m^{-2}$ for Gao et al. (2008) (Figs. S12 and S13). Interestingly, the simulated and reconstructed temperature responses are in better agreement at these timescales, in particular when using the Crowley and Unternman (2012) reconstruction, 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 0.25 °C. Changes in model simulations are smaller, lying between 0.05 and 0.1 °C 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

CPD
11, 2483–2555, 2015

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).

6.2 Framework for evaluation of climate model simulations: U_R and U_T statistics

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 done 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.

CPD
11, 2483–2555, 2015

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 and 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 versus 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 ensembles illustrates the degree of randomness in U_R and U_T statistics for in-

CPD
11, 2483–2555, 2015

dividual 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).

6.3 Detection and attribution

Figure 10 shows the results of the detection and attribution analysis using the multimodel ensemble mean, which is calculated as the mean of all model simulations described in Sect. 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.

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

CPD
11, 2483–2555, 2015

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 Eq. (2) (Sect. 3.5) is 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 undetectable, 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.

7 Discussion

In the light of the results presented in the Sects. 4 to 6, we discuss below each of the three questions raised in the introduction.

7.1 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 other cases. The consistency between simulations and observations is still generally more robust at hemispheric and global scales,

CPD
11, 2483–2555, 2015

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 Sect. 4.1). Thus, the published uncertainties for the reconstructions are in most cases too large to provide strong constrains 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 19th 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 (Supplement Sect. S3). Models also have some skill compared to a simple a priori estimate assuming no temperature change over the past millennium (Sect. 4.4). 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 (Sect. 6.2) that forced simulations are significantly closer to the reconstructions than unforced simulations in Northern Hemisphere regions. In particular, the Arctic region shows a particularly 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

CPD
11, 2483–2555, 2015

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 (Sect. 4.3), 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 (Sect. 4.1). Comparison of the series spectra (Sect. 4.2) 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 underestimation and overestimation of variance at different frequencies. 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

CPD
11, 2483–2555, 2015

the proxy data (e.g., Smerdon et al., 2010, 2015; Esper et al., 2012). 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 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 continental scale from the analyses performed herein, in contrast to some recent studies devoted to the ocean surface temperature (Laepple and Huybers, 2014a, b).

7.2 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 (Sect. 5.1), 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 (Sect. 5.2) and this is valid for most of the past millennium (Sect. 5.3). 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, 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.

CPD
11, 2483–2555, 2015

7.3 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 (Sect. 6.3). 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) (Sect. 6.2) 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 (Sect. 6.1). 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 (Sect. 6.3), which indicates that the magnitude of the simulated response to forcing in the Northern Hemisphere has the correct amplitude for smoothed time series. The

CPD
11, 2483–2555, 2015

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 better fit during the first 500 years (Fig. S18). This confirms earlier results obtained at hemispheric scale (Masson-Delmotte et al., 2013; Schurer et al., 2014).

The effect of external forcing is less clear for the Southern Hemisphere, where its influence is often not detected (Sect. 6.3). This is consistent with the lower correlation coefficients (Fig. S1) and weaker correlation statistics (U_R) there (Sect. 6.2). The models also seem to overestimate the response compared to the signal recorded in the Southern Hemisphere reconstructions (Sects. 6.2 and 6.3). 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 underestima-

CPD
11, 2483–2555, 2015

tion 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.

8 Conclusions and perspectives

The analysis of model simulations and PAGES 2k temperature reconstructions has allowed us to extend some of the 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

CPD
11, 2483–2555, 2015

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 continental scale seem less systematic than for some oceanic data (Laepple and Huybers, 2014a, b), 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

CPD
11, 2483–2555, 2015

disagreement between model simulations and Southern Hemisphere reconstructions (e.g., Neukom et al., 2014). Internal ocean variability can also have 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 output (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 detecting inconsistencies between model and palaeoclimate estimates.

Finally, one could also question the selection of the continental scale as 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 dy-

CPD
11, 2483–2555, 2015

namic interpretation (e.g., Luterbacher et al., 2004; Mann et al., 2009; Steiger et al., 2014; PAGES 2k Consortium, 2014).

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.

Team members

O. Bothe (The Ocean in the Earth System, Max Planck Institute for Meteorology, Hamburg, Germany), M. Evans (Department of Geology & ESSIC, University of Maryland, College Park, USA), L. Fernández Donado (Institute of Geoscience (UCM-CSIC), Department Astrophysics and Atmospheric Sciences, University Complutense Madrid, Spain), E. Garcia Bustamante (Department of Physics, University of Murcia, Murcia, Spain), J. Gergis (School of Earth Sciences, University of Melbourne, Australia), J. F. Gonzalez-Rouco (Institute of Geoscience (UCM-CSIC), Department Astrophysics and Atmospheric Sciences, University Complutense Madrid, Madrid, Spain), H. Goosse (ELIC/TECLIM, Université Catholique de Louvain, Louvain-la-Neuve, Belgium), G. Hegerl (School of GeoSciences, University of Edinburgh, Edinburgh, UK), A. Hind (Department of Physical Geography, Stockholm University, Sweden, Department of Mathematics, Stockholm University, Stockholm, Sweden), J. Jungclaus (The Ocean in the Earth System, Max Planck Institute for Meteorology, Hamburg, Germany), D. Kaufman (School of Earth Sciences & Environmental Sustainability, Northern Arizona University, Flagstaff, USA), F. Lehner (Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, USA, Climate and Environmental Physics and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland), N. McKay (School of Earth Sciences & Environmental Sustain-

CPD
11, 2483–2555, 2015

ability, Northern Arizona University, Flagstaff, USA), A. Moberg (Department of Physical Geography, Stockholm University, Stockholm, Sweden), C. C. Raible (Climate and Environmental Physics and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland), A. Schurer (School of GeoSciences, University of Edinburgh, Edinburgh, UK), F. Shi (Key Laboratory of Cenozoic Geology and Environment, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China), J. E. Smerdon (Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York, USA), L. von Gunten (PAGES International Project Office, Falkenplatz 16, 3012 Bern, Switzerland), S. Wagner (Institut for Coastal Research, Helmholtz Zentrum Geesthacht, Geesthacht, Germany), E. Warren (School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK), M. Widmann (School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK), P. Yiou (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), E. Zorita (Institut for Coastal Research, Helmholtz Zentrum Geesthacht, Geesthacht, Germany).

The Supplement related to this article is available online at doi:10.5194/cpd-11-2483-2015-supplement.

Author contributions. L. Fernández Donado, J. F. Gonzalez-Rouco, E. Garcia Bustamante, H. Goosse, J. Jungclaus organized the workshop at the origin of this paper. H. Goosse led the synthesis. O. Bothe, H. Goosse, G. Hegerl, A. Moberg, C. C. Raible, A. Schurer, S. Wagner, E. Zorita coordinated the writing. L. Fernández Donado, E. Garcia Bustamante, A. Hind, F. Lehner, N. McKay prepare the data sets and made them available to the whole group. O. Bothe, L. Fernández Donado, E. Garcia Bustamante, J. F. Gonzalez-Rouco, A. Hind, F. Lehner, N. McKay, A. Moberg, A. Schurer, S. Wagner, E. Warren, M. Widmann, E. Zorita performed the figures and their initial analysis. All authors contribute to the writing of the various sections and reviewed the manuscript.

CPD
11, 2483–2555, 2015

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. Goosse is Senior Research Associate 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). C. C. Raible and F. Lehner are supported by the Swiss National Science Foundation. P. Yiou is supported by the MILEX project of the Swedish Research Council. J. Gergis is funded by Australian Research Council Project DE130100668. O. Bothe was supported by LOCHMES (Leibniz-Society), PRIME-II (within DFG INTERDYNAMIK) and CliSAP. L. Fernández Donado was funded by a FPU grant: AP2009-4061. A. Moberg and A. Hind are supported by the Swedish Research Council grants B0334901 and C0592401.

The article processing charges for this open-access publication were covered by the Max Planck Society.

References

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

Allen, M. R. and Stott, P. A.: Estimating signal amplitudes in optimal fingerprinting, Part I: Theory, Clim. Dynam., 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, P. Natl. Acad. Sci. USA, 104, 3713–3718, doi:10.1073/pnas.0605064103, 2007.

CPD

11, 2483–2555, 2015

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, 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. Lett., 40, 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, 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. Sci., 35, 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 A, 54, 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, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, UK, New York, NY, USA, 867–952, 2013.

Bloomfield, P.: Fourier Analysis of Time Series: an Introduction, John Wiley and Sons, New York, 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. Climate, 26, 9399–9407, 2013.

CPD
11, 2483–2555, 2015

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, Astron. Astrophys., 202, 309–315, 1988.
Brohan, P., Allan, R., Freeman, E., Wheeler, D., Wilkinson, C., and 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. Climate, in press, doi: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. Climate, 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. Climate, 28, 2025–2043, 2015a.
Coats, S., Smerdon, J. E., Cook, B. I., and Seager, R.: Are simulated megadroughts in the North American Southwest forced?, J. Climate, 28, 124–142, doi:10.1175/JCLI-D-14-00071.1, 2015b.
Comboul, M., Emile-Geay, J., Hakim, G. J., and Evans, M. N.: Paleoclimatic sampling as a sensor placement problem, J. Climate, submitted, 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, doi:10.1126/sciadv.1400082, 2015.

CPD
11, 2483–2555, 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: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., and 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., and Wright, W. E.: Asian monsoon failure and megadrought during the last millennium, Science, 328, 486–489, doi:10.1126/science.1185188, 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 Syst. Sci. Data, 5, 187–197, doi:10.5194/essd-5-187-2013, 2013.

D'Arrigo, R., Wilson, R., and 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. Dynam., 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 Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim. Dynam., 40, 2123–2165, 2013.

CPD
11, 2483–2555, 2015

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: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 Climate 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, Quaternary Sci. 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 temperature response to external forcing in simulations and reconstructions of the last millennium, Clim. Past, 9, 393–421, 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?, Geophys. Res. Lett., 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, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, UK, New York, NY, USA, 741–866, 2013.

Flückiger, J., Dällenbach, A., Blunier, T., Stauffer, B., Stocker, T. F., Raynaud, D., and Barnola, J. M.: Variations in atmospheric N_2O concentration during abrupt climatic changes, Science, 285, 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 N_2O ice core record and its relationship with CH_4 and CO_2, Global Biogeochem. Cy., 16, 10-1–10-8, doi:10.1029/2001GB001417, 2002

CPD
11, 2483–2555, 2015

Franke, J., Frank, D., Raible, C. C., Esper, J., and Brönnimann, S.: Spectral biases in tree-ring climate proxies, Nature Climate 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. Climate, 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. Lett., 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?, Geophys. Res. Lett., 31, L06203, doi:10.1029/2003GL019140, 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, Quaternary Sci. 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, doi:10.5194/cp-2-99-2006, 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 Sci. Rev., 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. Dynam., 39, 2847–2866, doi:10.1007/s00382-012-1297-0, 2012b.
Hansen, J. and Sato, M.: Greenhouse gas growth rates, P. Natl. Acad. Sci. USA, 101, 16109–16114, doi:10.1073/pnas.0406982101, 2004.

CPD
11, 2483–2555, 2015

Hargreaves, J. C., Paul, A., Ohgaito, R., Abe-Ouchi, A., and Annan, J. D.: Are paleoclimate model ensembles consistent with the MARGO data synthesis?, Clim. 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. and Zwiers, F. W.: Use of models in detection and attribution of climate change, WIRES Climate Change, 2, 570–591, 2011.

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, 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. Climate, 20, 650–666, doi:10.1175/JCLI4011.1, 2007.

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

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

Hind, A., Moberg, A., and Sundberg, R.: Statistical framework for evaluation 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.

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, Climatic Change, 109, 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

CPD
11, 2483–2555, 2015

Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, UK, New York, NY, USA, 433–497, 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. Dynam., 20, 583–612, doi:10.1007/s00382-002-0296-y, 2003

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

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, Holocene, 19, 3–49, 2009.

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. Climate, 19, 3952–3972, 2006.

Jungclaus, J. H., Lorenz, S. J., Timmreck, C., Reick, C. H., Brovkin, V., Six, K., Segschneider, J., Giorgetta, M. A., Crowley, T. J., Pongratz, J., Krivova, N. A., Vieira, L. E., Solanki, S. K., Klocke, D., Botzet, M., Esch, M., Gayler, V., Haak, H., Raddatz, T. J., Roeckner, E., Schnur, R., Widmann, H., Claussen, M., Stevens, B., and Marotzke, J.: Climate and carbon-cycle variability over the last millennium, Clim. 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, 2201–2213, doi:10.5194/cp-10-2201-2014, 2014.

Kaplan, J. O., Krumhardt, K. M., Ellis, E. C., Ruddiman, W. F., Lemmen, C., and Goldewijk, K. K.: Holocene carbon emissions as a result of anthropogenic land cover change, Holocene, 21, 775–779, doi:10.1177/0959683610386983, 2011.

CPD
11, 2483–2555, 2015

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. Lett., 32, L20708, doi:10.1029/2005GL024233, 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, 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, Climatic 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, 2528–2534, doi:10.1002/2014GL059345, 2014a.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: large model–data differences at decadal and longer periods, P. Natl. Acad. Sci. USA, 11, 16682–16687, doi:10.1073/pnas.1412077111, 2014b.

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., 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, 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. Climate, 26, 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, Quaternary Sci. Rev., 45, 85–94, 2012.

Lehner, F., Born, A., Raible, C. C., and Stocker, T. F.: Amplified inception of European Little Ice Age by sea ice–ocean–atmosphere feedbacks, J. Climate, 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 Syst. Dynam. Discuss., 6, 351–406, doi:10.5194/esdd-6-351-2015, 2015.

CPD
11, 2483–2555, 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, 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 CO_2, CH_4 and N_2O ice core records extended to 2000 years BP, Geophys. Res. Lett., 33, L14810, 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. Lett., 22, 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. Lett., 26, 759–762, 1999.

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.

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

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, Mon. Weather Rev., 139, 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, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, UK, New York, NY, USA, 383–464, 2013.

CPD
11, 2483–2555, 2015

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. Geosci., accepted, 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.: 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. and Brattström, G.: Prediction intervals for climate reconstructions with autocorrelated noise – an analysis of ordinary least squares and measurement error methods, Palaeogeogr. Palaeocl., 308, 313–329, 2011.

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

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, 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.,

CPD
11, 2483–2555, 2015

Zinke, J., and Frank, D.: Inter-hemispheric temperature variability over the last millennium, Nature Climate Change, 4, 362–367, 2014.
Ortega, P., Lehner, F., Casado, M., Swingedouw, D., Masson-Delmotte, V., Yiou, P., and 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., and Zorita, E.: Continental-scale temperature variability during the last two millennia, Nat. Geosci., 6, 339–346 doi:10.1038/NGEO1797, 2013.
PAGES 2k Consortium (Primary Authors: Anchukaitis, K., Büentgen, U., Emile-Geay, J., Evans, M. N., Goosse, H., Kaufman, D., Luterbacher, J., Smerdon, J., Tingley, M., and von Gunten, L.): A community-driven framework for climate reconstructions, EOS T. Am. Geophys. Un., 95, 361–362, doi:10.1002/2014EO400001, 2014.
Percival, D. B. and Walden, A. T.: Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques, Cambridge University Press, Cambridge, UK, 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. Climate, 26, 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, Global Biogeochem. Cy., 22, GB3018, 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, Global Biogeochem. Cy., 23, GB4001, doi:10.1029/2009GB003488, 2009.

CPD
11, 2483–2555, 2015

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., Casty, C., Luterbacher, J., Pauling, A., Esper, J., Frank, D. C., Büntgen, U., Roesch, A. C., Tschuck, P., Wild, M., Vidale, P.-L., Schär, C., and Wanner, H.: Climate variability – observations, reconstructions, and model simulations for the Atlantic-European and Alpine region from 1500–2100 AD, Climatic 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, 852–863, 2013.

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

CPD
11, 2483–2555, 2015

Journal of Advances in Modeling Earth Systems, 6, 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. Geosci., 7, 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., and Xiao, X.: Northern Hemisphere temperature reconstruction during the last millennium using multiple annual proxies, Clim. Res., 56, 231–244, 2013.
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.: 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., 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. Climate, 23, 4856–4880, 2010.
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. Climate, 28, 3212–3231, 2015a.
Smerdon, J. E., Coats, S., and Ault, T. R.: Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the common era, Clim. Dynam., doi:10.1007/s00382-015-2684-0, online first, 2015b.
Steiger, N. J., Hakim, G., Steig, E. J., Battisti, D. S., and Roe, G. H.: Assimilation of time-averaged pseudoproxies for climate, J. Climate, 27, 426–441, 2014.
Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, Geophys. Res. Lett., 36, L19704, 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.

CPD
11, 2483–2555, 2015

Stothers, R. B.: The great Tambora eruptions in 1815 and its aftermath, Science, 224, 1191–1198, 1984.
Sundberg, R., Moberg, A., and 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, B. Am. Meteorol. Soc., 93, 485–498, 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. Dynam., 28, 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, Quaternary Sci. 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, P. 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, Y. M., Lean, J. L., and Sheeley Jr, N. R.: Modeling the sun's magnetic field and irradiance since 1713, Astrophys. J., 625, 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, Clim. Past, 6, 627–644, doi:10.5194/cp-6-627-2010, 2010.

CPD
11, 2483–2555, 2015

Wilmes, S. B., Raible, C. C., and Stocker, T. F.: Climate variability of the mid- and high-latitudes of the Southern Hemisphere in ensemble simulations from 1500 to 2000 AD, Clim. Past, 8, 373–390, doi:10.5194/cp-8-373-2012, 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, B. Am. Meteorol. 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, 11, 939–958, doi:10.5194/cp-11-939-2015, 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, doi:10.5194/tc-7-451-2013, 2013.

CPD
11, 2483–2555, 2015

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° × 1.25°, L26 (atm) nominal 1°, L60 (ocn)	288 × 192, L26 (atm) 320 × 384, L60 (ocn)	10	20	30, 31, 32	40	50	60	Landrum et al. (2013)	500
CESM1	1	0.9° × 1.25°, L26 (atm) nominal 1°, L60 (ocn)	288 × 192, L26 (atm) 320 × 384, L60 (ocn)	11	20	30, 31, 32	40	50	1990 CE	Lehner et al. (2015)	465
CSIRO-Mk3L-1-2	1	5.63° × 3.21°, L18 (atm) 2.81° × 1.61°, L21 (ocn)	64 × 56, L18 (atm) 128 × 112, L21 (ocn)	12	21	30, 31, 32	none	none	60	Phipps et al. (2013)	1150
GISS-E2-R	3	2° × 2.5°, L40 (atm) 1° × 1.25°, L32 (ocn)	144 × 90, L40 (atm) 288 × 180, L32 (ocn)	12	21, 20	30, 31, 32	40	50, 51	60	Schmidt et al. (2014b)	1162
HadCM3	1	3.75° × 2.46°, L19 (atm) 1.25° × 1.25°, L20 (ocn)	96 × 73, L19 (atm) 288 × 144, L20 (ocn)	12	21	30, 33, 32	41	51	60	Schurer et al. (2013)	1199
IPSL-CM5A-LR	1	3.75° × 1.88°, L17 (atm) 1.98° × 1.21°, L32 (ocn)	96 × 96, L17 (atm) 182 × 149, L32 (ocn)	10	22	30, 31, 32	none	none	60	Dufresne et al. (2013)	1004
MPI-ESM-P	1	1.84° × 1.84°, L47 (atm) nominal 1.5°, L40 (ocn)	196 × 98, L47 (atm) 256 × 220, L40 (ocn)	10	21	30, 31, 32	40	52	60	Jungclaus et al. (2014)	1155
ECHAM5/MPIOM (COSMOS)	E1:5	3.75° × 3.75°, L19 (atm) nominal 3°, L40 (ocn)	96 × 48, L19 (atm) 120 × 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).

CPD
11, 2483–2555, 2015

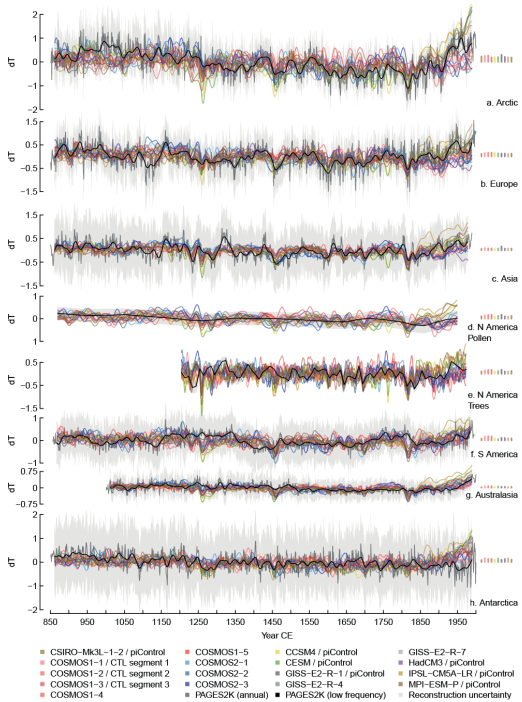

Figure 1. Series of simulated temperatures and PAGES 2k reconstructions for the seven continent-scale regions. The reconstructions are shown at their original resolution and after a smoothing using a 23 year Hamming filter, except for the North American reconstructions. Only the smoothed series are shown 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 full length of temporal overlap between simulations and reconstruction. Note the different scales in the *y* axis of the various regions.

CPD
11, 2483–2555, 2015

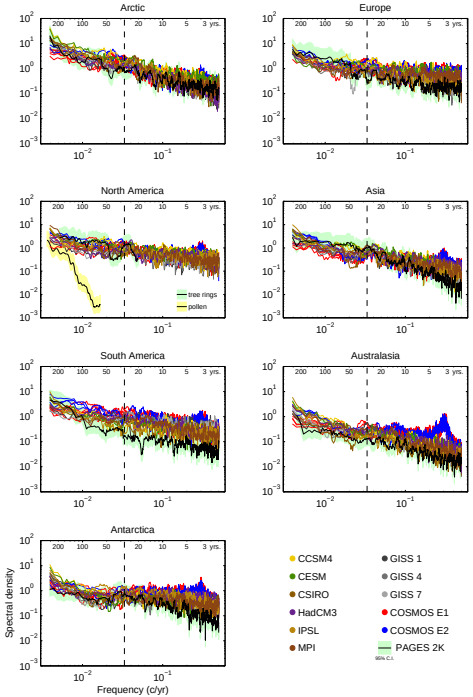

Figure 2. Spectral densities for all simulations and reconstructions computed with the multi-taper method, with the time-bandwidth product set to 4. Units are temperature variance (°C^2 or K^2) per frequency (c year^{-1}). Spectra are calculated using all existing data in the period 850–2005, with the long-term averages subtracted, and are plotted for frequencies larger than the analysis bandwidth and smaller than 0.5 times the sampling rate. Reconstruction spectra are illustrated with their 95 % confidence intervals, while model spectra are shown with single coloured lines. Dashed vertical lines drawn at the frequency 1/30 yr^{-1} denote the limit for frequencies of relevance for analyses made at the 15 year resolution, or with a 23-point Hamming window, as in many other analyses in this study.

CPD
11, 2483–2555, 2015

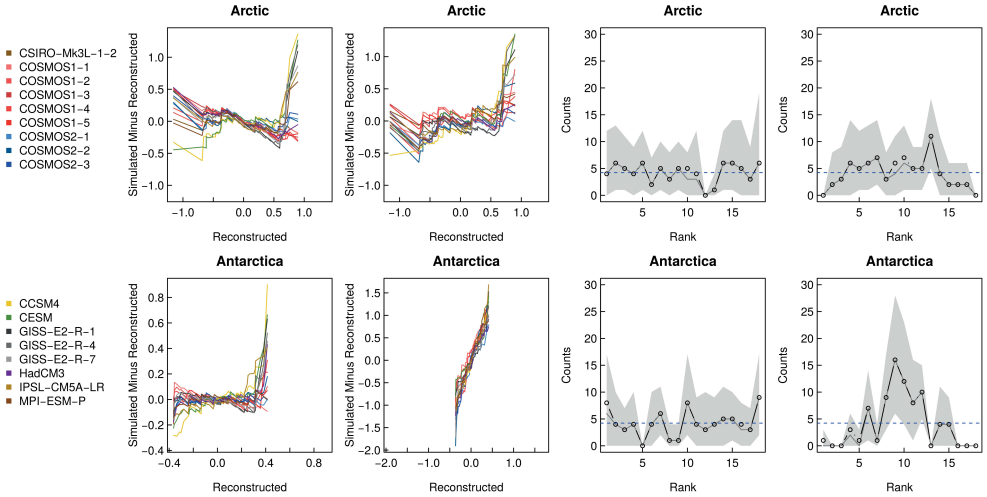

Figure 3. Climatological consistency (first two columns): residual quantile-quantile plots for the full period; and probabilistic consistency (last 2 columns): rank counts for the full period. The top row is for the Arctic, and the bottom row is for Antarctica. For both the climatological and probabilistic consistency, the computations are obtained by neglecting the uncertainties (left plot) and using the uncertainties provided with the original reconstructions (right plot).

CPD
11, 2483–2555, 2015

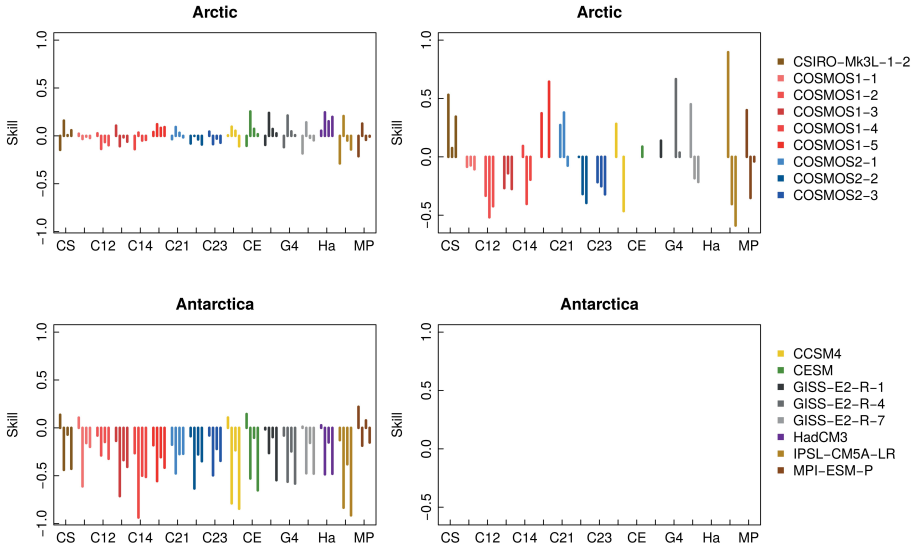

Figure 4. Skill metric for the individual models for all periods (from left to right: 850–1350, 1350–1850, 850–1850, 850–2000). Top row for the Arctic, bottom for Antarctica. The computations assume no uncertainties (left plot) and uncertainties provided with the original reconstructions (right plot).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

CPD
11, 2483–2555, 2015

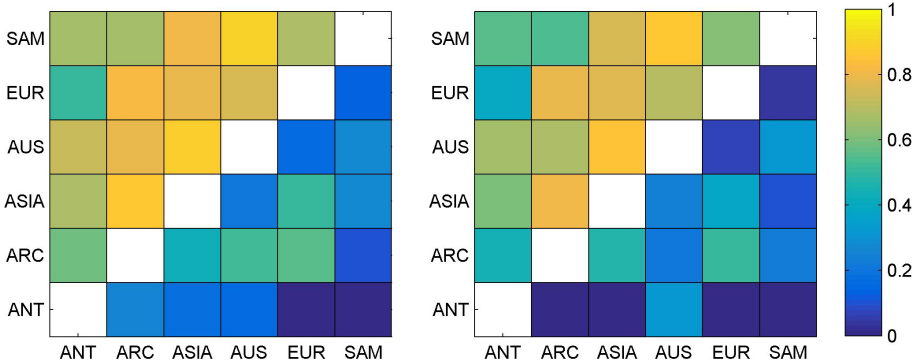

Figure 5. Correlations among the PAGES 2k regions for detrended simulated and reconstructed time series filtered using a 23 year Hamming filter. Left-hand panel: forced simulation with MPI-ESM (upper triangle) PAGES 2k reconstructions (lower triangle) for 1012–1978 CE. Right-hand panel: forced simulation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

CPD
11, 2483–2555, 2015

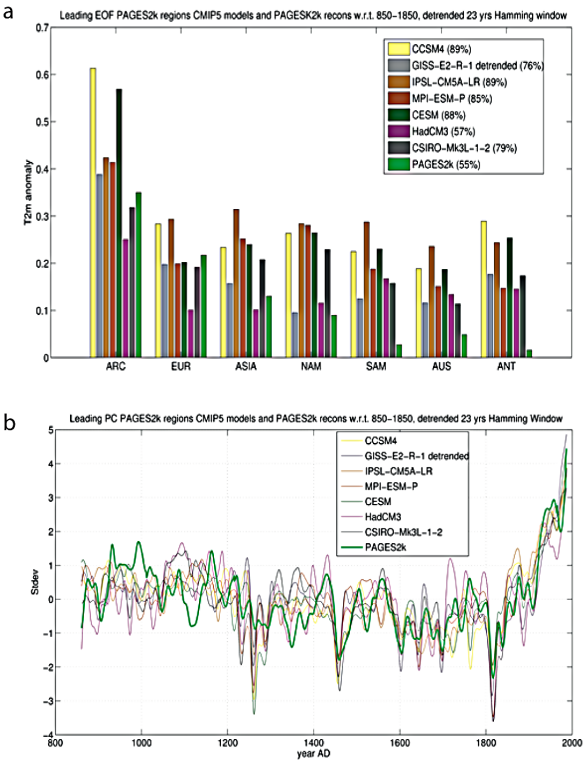

Figure 6. (a) EOFs of the near-surface temperature simulated by each CMIP5/PMIP3 model and reconstructions over the period 850–2004 CE. The eigenvectors are based on the covariance matrix with respect to temperature anomalies for the period 850–1850 CE. **(b)** Time series of the principal components 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.

CPD
11, 2483–2555, 2015

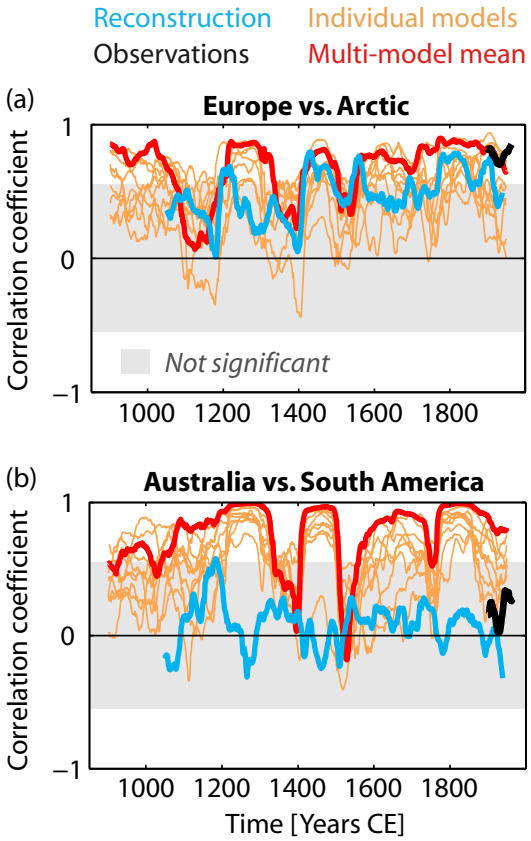

Figure 7. 100 year moving Tukey window correlations between selected PAGES 2k regions for the PAGES 2k reconstructions (blue) and PMIP3 models (8 models in orange, multi-model mean in red) and observations from HadCRUT4 (Morice et al., 2012, black). Each 100 year segment is linearly detrended beforehand. Grey shading illustrates correlations that are not significant at the 5 % level.

CPD
11, 2483–2555, 2015

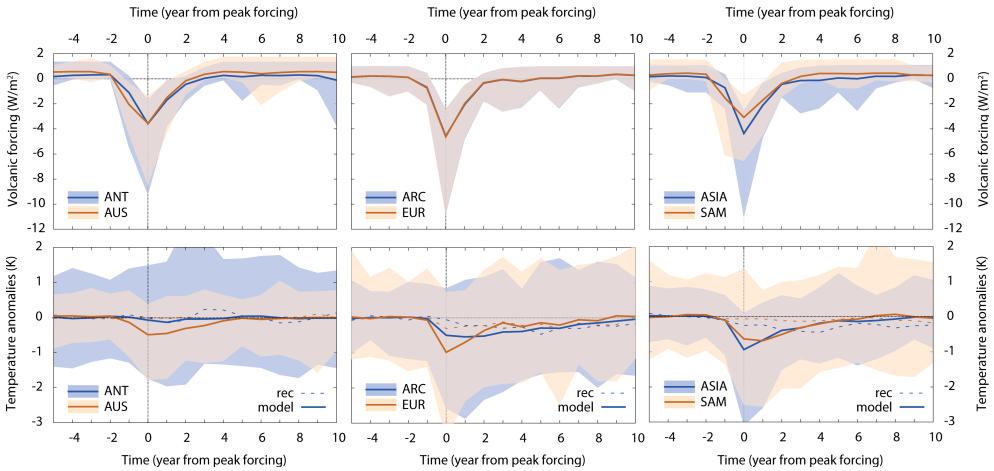

Figure 8. Superposed composites of volcanic forcing and temperature responses during selected periods when peak negative forcing in the Crowley and Unterman (2012) reconstruction are aligned. The composite is produced by selecting the 12 strongest volcanic events, starting 5 years before the date of the peak eruption and ending 10 years after the event. Top panels show the average volcanic forcing for each PAGES 2k region (lines) and the range of values attained in the 12 episodes (shading). Bottom panels indicate the reconstructed (dashed lines) and simulated (solid) temperature composite for the same events. The range of simulated temperatures is also indicated with the colour shading.

CPD
11, 2483–2555, 2015

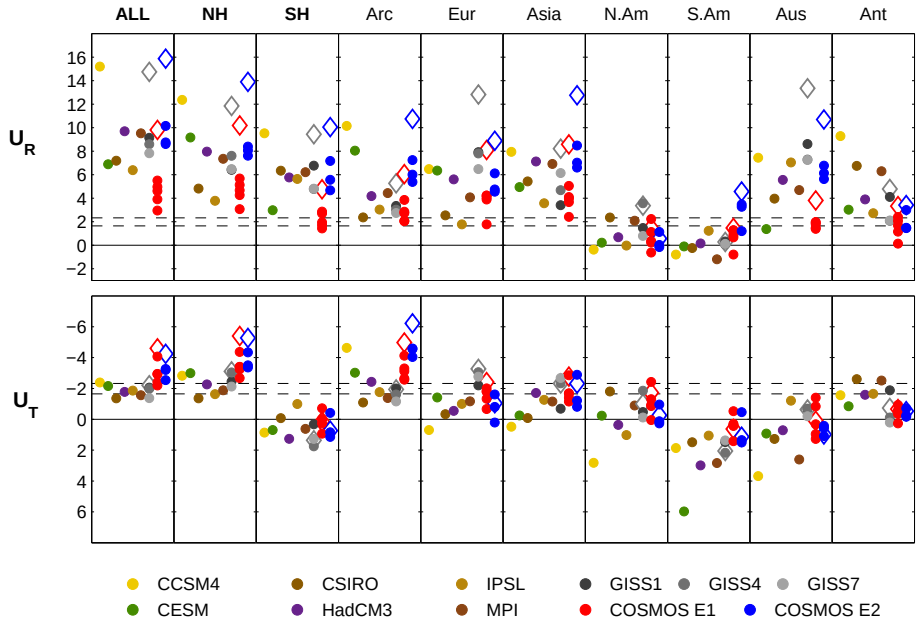

Figure 9. U_R and U_T statistics for PAGES 2k regions in the period 861–1850 CE. Coloured dots: individual simulations. Diamonds: ensemble-mean results for COSMOS and GISS models. Positive U_R indicates that simulations and reconstructions share an effect of temporal changes in external forcings. Negative U_T indicates that a forced simulation is closer to the observed temperature variations than its own control simulation. Dashed lines show one-sided 5 and 1 % significance levels. Note the reversed vertical axis in the U_T graphs.

CPD
11, 2483–2555, 2015

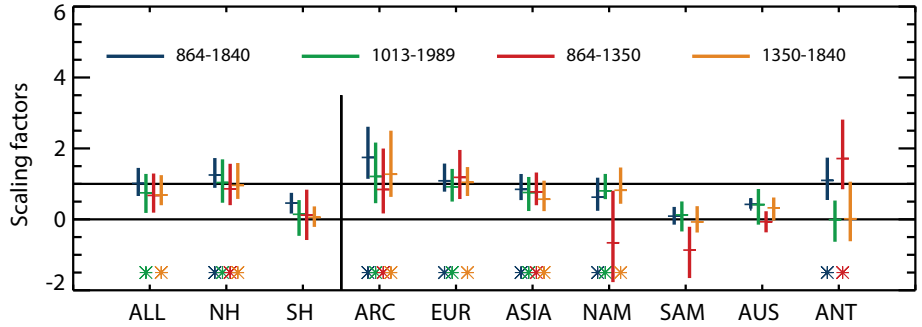

Figure 10. Detection and attribution results for PAGES 2k regions. For each region scaling ranges are shown for four different time periods: Northern Hemisphere (NH), Southern Hemisphere (SH) and global (ALL). The regression was carried out on the combined data from all the applicable regions. An asterisk 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.

CPD
11, 2483–2555, 2015