Clim. Past Discuss., 9, 3789–3824, 2013
www.clim-past-discuss.net/9/3789/2013/
doi:10.5194/cpd-9-3789-2013

Climate of the Past
Discussions
Open Access

Consistency of the multi-model CMIP5/PMIP3-past1000 ensemble

O. Bothe[1,2,*], **J. H. Jungclaus**[1], **and D. Zanchettin**[1]

[1]Max Planck Institute for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany
[2]University of Hamburg, KlimaCampus Hamburg, Hamburg, Germany
[*]now at: Leibniz Institute of Atmospheric Physics at the University of Rostock, Kühlungsborn, Germany

Received: 24 June 2013 – Accepted: 29 June 2013 – Published: 9 July 2013

Correspondence to: O. Bothe (ol.bothe@gmail.com)

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

Abstract

We present an assessment of the probabilistic and climatological consistency of the CMIP5/PMIP3 ensemble simulations for the last millennium relative to proxy-based reconstructions under the paradigm of a statistically indistinguishable ensemble. We evaluate whether simulations and reconstructions are compatible realizations of the unknown past climate evolution. A lack of consistency is diagnosed in surface air temperature data for the Pacific, European and North Atlantic regions. On the other hand, indications are found that temperature signals partially agree in the western tropical Pacific, the subtropical North Pacific and the South Atlantic. Deviations from consistency may change between sub-periods, and they may include pronounced opposite biases in different sub-periods. These distributional inconsistencies originate mainly from differences in multi-centennial to millennial trends. Since the data uncertainties are only weakly constrained, the frequent over-dispersive distributional relations prevent the formal rejection of consistency of the simulation ensemble.

1 Introduction

The fifth phase of the coupled model intercomparison project (CMIP5, Taylor et al., 2012) incorporates, for the first time, paleoclimate simulations in its suite of numerical experiments. The last 1000 yr of the pre-industrial period are the most recent key-period identified by the Paleoclimate Modelling Intercomparison Project Phase III (PMIP3, Braconnot et al., 2012). In contrast to the traditional time-slice simulations for specific periods of the past (e.g. Last Glacial Maximum), the PMIP3 "past1000" experiments are transient simulations covering 850 to 1850 AD with time-varying estimates for external drivers, such as orbital, solar, volcanic and land-use climate forcings (Schmidt et al., 2011). The past1000-ensemble bridges a gap between the unperturbed control simulations and the historical simulations for the last 150 yr. It provides simulated estimates of a climate only slightly different from today. Since the ensemble allows

CPD
9, 3789–3824, 2013

for detailed comparisons with climate reconstructions it assists in improving our understanding of past climate forcings and naturally-forced climate variability and, in turn, in fingerprinting anthropogenic climate change (Hegerl et al., 2007; Sundberg, 2012; Schmidt et al., 2013). Assessing the quality of our simulations against paleoclimate estimates provides essential testbeds for our climate models (e.g. Schmidt et al., 2013).

Commonly, validation considers how accurately a simulated data set agrees to the observational data in terms of matching patterns (e.g. Taylor, 2001). Comparison of simulations and reconstructions implicitly interprets both as representations of the same past. Based on this, their agreement may be taken as validation of the model and their disagreement may highlight model deficiencies. However, we have to take into account the considerable uncertainties in the reconstructions. Thus, we propose that it is appropriate in the past1000-context to assess the consistency of the simulations applying methods from weather-forecast verification following, e.g. Annan and Hargreaves (2010) and Marzban et al. (2011) prior to any subjective comparison. This means that we have to ask whether simulations and reconstructions represent compatible realizations of the unknown past climate in terms of the distribution from which its trajectory samples under the operating forcings (e.g. Hargreaves et al., 2011; Bothe et al., 2013). Such consistency provides confidence that simulations and reconstructions indeed describe the same property and include comparable amounts of forced and internal variability. The paradigm of a statistically indistinguishable ensemble offers a theoretical basis to evaluate the consistency of simulation ensembles with reconstructions. We use this framework to assess the ensemble consistency of the past1000 multi-model-ensemble with the global temperature field reconstruction by Mann et al. (2009) and two regional area-averaged temperature reconstructions for Central Europe (Dobrovolný et al., 2010) and Southwestern North America (Wahl and Smerdon, 2012). We interpret the past1000-ensemble in terms of a probabilistic ensemble of realizations of climate variability and test whether it reliably represents the reconstructed distribution including the reconstruction uncertainty.

CPD
9, 3789–3824, 2013

The current study extends our previous work (Bothe et al., 2013), which tested the consistency of the PMIP3-compliant ensemble of simulations for the last millennium performed with the COSMOS version of the MPI Earth System Model (MPI-ESM) developed at the Max Planck Institute for Meteorology (COSMOS-Mill ensemble, Jungclaus et al., 2010). The ensemble spans a number of forcing and initial conditions. We found that the COSMOS-Mill ensemble commonly lacks consistency with a set of reconstructions for Northern Hemisphere mean temperature and with the global temperature field reconstruction by Mann et al. (2009). However, its representations of Central European annual mean temperature are consistent with the reconstruction by Dobrovolný et al. (2010).

The PMIP3-past1000 multi-model-ensemble allows considering the consistency of our paleoclimate-simulations with reconstructions not only under initial and forcing condition uncertainties (like for the COSMOS-Mill ensemble) but especially under the different parametric choices in different models and different structural uncertainties in the models (Mauritsen et al., 2012; Tebaldi and Knutti, 2007). The models contributing to CMIP5 and PMIP3 (Taylor et al., 2012) generally represent an improvement over the previous generation of models which contributed to CMIP3 (Meehl et al., 2007). In particular, some models (including MPI-ESM) provide paleo-simulations at the same resolution as the historical and the future scenario simulations. Schmidt et al. (2013) emphasise the importance of such a setup for paleo-simulations to be useful in assessing the quality of simulations of the 20th century and of future climate projections. However, in contrast to the COSMOS-Mill ensemble, none of the past1000-simulations were performed including calculations of a carbon cycle. We consider the multi-model analysis to clarify the consistency of simulations under the parametric differences between the models and the common or distinct structural uncertainties of the models (e.g. Sanderson and Knutti, 2012). Therefore, we do not expect a priori increased consistency compared to our earlier results (Bothe et al., 2013).

Section 2 gives details on the methodological approach and the employed data before Sect. 3 presents results on the consistency of the past1000-ensemble with the reconstructions. In Sect. 4, we identify sources for the found (lack of) consistency. Short concluding remarks close the manuscript.

2 Methods and data

2.1 Methods

To build confidence in a simulation ensemble we may either consider the accuracy of its members in reproducing a given (observed) target (e.g. following Taylor, 2001) or assess its statistical consistency with a target data set (see, e.g. Marzban et al., 2011). The evaluation of ensemble-consistency follows the paradigm of a statistically indistinguishable ensemble (for a more detailed discussion of the methods see, e.g. Bothe et al., 2013). The underlying null-hypothesis is that the verification target and the simulations are samples from a larger distribution and therefore exchangeable (Annan and Hargreaves, 2010; Rougier et al., 2012). In the paleoclimate context, climate reconstructions are our best estimate of an observed target.

We analyse the ensemble-consistency based on two points of view. Firstly, probabilistic consistency considers the multivariate distribution of ensemble and verification data, and, secondly, climatological consistency considers the climatological distribution of the individual simulations (e.g. Johnson and Bowler, 2009; Marzban et al., 2011; Wilks, 2011).

The probabilistic evaluation addresses how the frequencies of occurrence of the ensemble data compare to those of the verification data. It allows assessing the ensemble variance but also detecting biases. The climatological evaluation analyses the climatological variance and the biases within the individual ensemble members in relation to that of the verification data.

We can assess the probabilistic component of consistency for an ensemble by sorting and ranking the target against the ensemble data (e.g. Anderson, 1996; Jolliffe and Primo, 2008; Annan and Hargreaves, 2010; Marzban et al., 2011; Hargreaves et al., 2011). Counts of the calculated ranks are displayed as histograms following Anderson (1996). Under the null hypothesis of exchangeability (i.e. indistinguishability) of the distributions, a histogram should be flat since frequencies of observed and ensemble estimated data agree for a consistent ensemble (Murphy, 1973). The criterion of flatness does not state that the ensemble indeed is consistent (as discussed by, e.g. Hamill, 2001; Marzban et al., 2011) but it is a necessary condition for our ensemble to be a reliable representation relative to the chosen verification.

Marzban et al. (2011) emphasize the climatological consistency. They propose to evaluate it by plotting the difference between the simulated and the target quantiles against the target quantiles. For a consistent simulation, such residual quantile-quantile plots should display a flat outcome at zero. Residual quantile-quantile (r-q-q) plots ease the interpretation compared to conventional quantile-quantile plots. Marzban et al. (2011) and Bothe et al. (2013) provide more details on the advantages of the r-q-q plots.

Residual quantiles and rank counts provide an easily understandable visualization of deviations of the ensemble relative to the verification data. In r-q-q plots, biases of the ensemble data are seen as displacements from the zero line. A positive slope in the residual quantiles highlights an overestimation of the difference of the quantiles to the mean (i.e. the variance) compared to the target quantiles. Such a too wide data set is called over-dispersive. On the other hand, a negative slope highlights an underestimation of the variance, a too narrow data set, to which we refer as under-dispersive.

In rank histograms, dome-shapes (U-shapes) indicate too wide (too narrow) probabilistic distributions, i.e. verification data are more often close to (distant from) the mean of the distribution compared to the simulation ensemble. Positive (negative) slopes represent negative (positive) ensemble-biases, i.e. the target data over-populate high (low) ranks.

We use the χ^2 goodness-of-fit test to test for the consistency of a rank count with the uniform, i.e. flat, null hypothesis. Jolliffe and Primo (2008) provide a decomposition of the test to further consider individual deviations from the expected flat outcome.

CPD
9, 3789–3824, 2013

These are, among others, bias and spread deviations. Goodness-of-fit statistics are presented for these two single-deviation tests and the full test and discussed in terms of their p values with respect to the upper 90 % critical values (for single deviation tests it equals 2.706, for the full test with 8 degree of freedom of the χ^2 distribution it equals 13.362).

The analyses require the target to provide an accurate representation of the past climate trajectory, a condition that is hardly met in paleoclimate studies due to the associated uncertainties (e.g. Wilson et al., 2007; Bradley, 2011; Randall et al., 2007; Schmidt et al., 2011, 2013). Otherwise, we have to include an uncertainty estimate in the ensemble data (Anderson, 1996). The reconstruction targets uncertainty estimates are used to inflate the simulated data.

Analyses under the paradigm of statistical indistinguishability require special care if we use them in the context of paleoclimatology. Any simulated or reconstructed time series over the last 1000 yr includes components of forced and internal variability. If we assume that our estimates of past forcings are approximately correct, there should be a common forced signal in simulations and reconstructions. However, the reconstruction data uncertainties are possibly a lower bound for the disagreement between simulations and their validation target (Schmidt et al., 2013). Our analysis identifies whether the variability, forced and internal, originates from distributions that are similar enough to be indistinguishable. In this case, we would state that the reconstruction and the simulation ensemble are consistent. If the variability of the ensemble data deviates significantly from that of the target, our approach identifies inconsistencies. The approach can also be used to highlight in which period the long-term signals do not agree between the reconstruction and the simulation ensemble. Arising lack of consistency in terms of the distributional characteristics indicates that the simulations and the reconstruction provide different representations of the past climate. Such deviations are informative as they suggest a need for model and reconstruction-method improvements. However, they also limit the validity of conclusions on past climate variability, the climate response to past forcings or the anthropogenic fingerprint.

CPD
9, 3789–3824, 2013

The assessment of consistency reduces, in principle, the subjectivity associated to the comparison of simulations and reconstructions (compare Bothe et al., 2013). However, the large uncertainties require re-considering the importance of distributional deviations (compare Hargreaves et al., 2011; Annan et al., 2011). Over-dispersion does not necessarily question the overall reliability of the ensemble (see Hargreaves et al., 2011; Annan et al., 2011). On the other hand, if a simulation ensemble is found to be too narrow or biased under uncertainty, further simple comparison studies between the ensemble and the reconstruction may be misleading on the considered scales. We have to consider the suggested lack of consistency in subsequent research, but we may also conclude that, under the present uncertainties, comparison of simulated and reconstructed estimates is not informative. Note, consistency (i.e. exchangeability) and agreement (i.e. accuracy of the temporal patterns) may differ regionally; inconsistent regions can agree in the signal and regions lacking common signals can be consistent.

2.2 Data

If we want to achieve a robust evaluation of the consistency of a simulation ensemble, we have to consider, ideally, more than one data set and more than one parameter, not least because of the prominent uncertainties in climate reconstructions. However, the global temperature field reconstruction by Mann et al. (2009) is the only data that allows for a globally coherent evaluation of a climate parameter for the last millennium. It consists of decadally smoothed data. We further employ area-averaged temperature reconstructions focused on showing two data-sets for the last 500 yr for temperature in Central Europe (Dobrovolný et al., 2010) and Southwestern North America (Wahl and Smerdon, 2012). These data sets serve as examples for area-averaged reconstructions. The focus is motivated by the assumption that reconstructions and forcing data are generally more reliable on this period (compare Bothe et al., 2013).

The past1000-simulations available from the CMIP5 database were performed with the following models (see Table 1): BCC, CCSM4 (Landrum et al., 2013), FGOALS (Zhou et al., 2011), GISS-R (two realizations, http://data.giss.nasa.gov/modelE/ar5/),

CPD
9, 3789–3824, 2013

IPSL-CM5A-LR (Hourdin et al., 2012) and the current generation of MPI-ESM (e.g. including an updated version of the atmospheric general circulation component ECHAM6, Giorgetta et al., 2013; Jungclaus et al., 2013). The CSIRO PMIP3-only simulation is included in our ensemble (different from Phipps et al., 2012). We do not consider the further members of the GISS-R-ensemble (R21, R22). These would overemphasize the influence of GISS-E2-R on the ensemble measures.

We exclude the simulation with MIROC-ESM (by the Atmosphere and Ocean Research Institute at the University of Tokyo, the National Institute for Environmental Studies, and Japan Agency for Marine–Earth Science and Technology) since it shows a problematic long-term drift (A. Abe-Ouchi, personal communication, 2012). On the other hand, a simple correction is performed for the drift of the GISS-R simulations (G. A. Schmidt, personal communication, 2012, see also Schmidt et al., 2012): we subtract a lowess fit (influence of about 600 yr) to the GISS-R pre-industrial control-run (piControl) from the annually resolved data of interest, i.e. the grid-point data for the field evaluation and the relevant time series for the area averaged assessment.

We generally use non-overlapping decadal means for simulations and reconstructions in the commonly included period 1000–1849 CE and anomalies relative to this period. For the data from Mann et al. (2009) we choose the central date for each decade (i.e. 1844 for the 1840s) since the data are originally decadally smoothed. Results change slightly but conclusions are the same when we employ non-overlapping decadal means for this reconstruction. We also employ the three sub-periods 1000s–1270s, 1280s–1550s, 1560s–1830s to evaluate how consistency may change over time. Inclusion of climates strongly biased from our reference period (e.g. the industrial period) would complicate our assessment of consistency focused on paleoclimates. The different grids of the simulations require interpolating the data onto a common T21-grid ($\sim 5^{\circ}$).

The global field reconstruction further allows evaluating the consistency of approximations of major climate indices which are commonly interpreted to present low-dimensional descriptors of the climate system (compare, e.g. Tsonis et al., 2011; Dima and Lohmann, 2007). Our approach is, as well for these data, a step beyond the pure "by eye" approaches of reconstruction-simulation assessment. We construct two indices as field averages over the Pacific (150° E–140° W, 24–53° N) and Atlantic (74° W–0° E, 2–53° N) domains. Higher latitudes are excluded to avoid effects from sea-ice variability. Simulated indices are calculated from surface air temperatures, in contrast to the common definition via sea-surface or upper ocean temperatures. This appears justified since the reconstruction is a hybrid representation of sea-surface and near-surface air temperature (compare Brohan et al., 2006; Mann et al., 2009) with only a minority of underlying proxies being of marine origin. The indices are denoted by AMO (Atlantic Multidecadal Oscillation) and PDO (Pacific Decadal Oscillation), although our definitions differ from the convention (e.g. Zanchettin et al., 2013, and references therein). We do not preprocess the input data and do not standardize the series. The indices accumulate globally- and regionally-forced as well as potential internal signals.

We have to include an uncertainty estimate in our analyses by inflating the simulation ensemble (compare Anderson, 1996). These are generally randomly selected from a Gaussian distribution with zero mean. For the regional reconstructions, the standard deviations are the reported standard errors, while, for the constructed indices, we use a standard deviation of 0.2 K which approximates the estimates given for similar indices by Mann et al. (2009). For the field reconstruction, we follow Bothe et al. (2013) and take the largest standard error ($\sigma \approx 0.1729$) reported for the Northern Hemisphere mean temperature series of Mann et al. (2009) as a reasonable uncertainty estimate for the field data.

We observe that neglecting uncertainties in the reconstruction can lead to pronounced differences in our inferences about the probabilistic and climatological consistency of the ensemble. That is, the ensemble may appear under-dispersive or even consistent excluding the uncertainties, although it is found to be over-dispersive if they are considered.

Our knowledge of past forcings is rather weakly constrained as seen in comparisons of the available reconstructions for land-use, total solar irradiance and volcanic eruptions as compiled by Schmidt et al. (2011, see also discussion by Schmidt et al., 2013). For the employed simulations, differences are especially noted in the volcanic forcing, which in turn implies that they mostly influence the annual time-scale and pentad data. We will recur to this point in our discussion of origins of (lack of) consistency in Sect. 4.

Consistency in our setting depends on the reference time. Due to the way the reconstructions are produced, it is in principle advisable to center all data on the calibration period of the reconstruction (J. Smerdon, personal communication, 2012) since this time is the reference for the calculation of uncertainties. We instead center our data over the full studied period, and thereby shift the focus from the complete comparability over pre-industrial and industrial times to the comparability of the variability over the pre-industrial time only.

We stress that our results neither allow to rank the various simulators against one another nor to decide whether individual simulations or even the ensemble mean rather than the reconstruction are more representative of past climate variability.

3 Results

3.1 Global field consistency

Figure 1 gives a first impression of the probabilistic consistency of the past1000-ensemble with the global temperature field reconstruction by Mann et al. (2009). We display the p values of tests for a uniform outcome of the rank counts at every grid-point. The goodness-of-fit test leads to the rejection of the null hypothesis of a uniform outcome at grid-points for a p value larger than 0.9 (red in Fig. 1). Thus, the analysis shows a lack of consistency for large areas of the globe for the full period (Fig. 1a). In contrast, the European Arctic is the only spatially extended area for which rank counts deviate significantly from uniformity for an arbitrary shorter period (Fig. 1b). Possible consistency is diagnosed elsewhere. If we test for individual deviations of bias

CPD
9, 3789–3824, 2013

or spread, at least one of them is significant over much of the globe for both, the full and the shorter period (Fig. 1c, and d).

This general impression has to be complemented by a detailed look at individual locations to identify the character of the inconsistencies. Considering a sample of grid-points and three different sub-periods, residual quantile distributions display various different structures when evaluating the climatological consistency of the ensemble (Fig. 2, left panels, extended sample in the Supplement). A too wide simulated distribution arises as a common feature due to very strong overestimation of cold anomalies in the simulated data and notable overestimation of positive anomalies, i.e. we see a positive slope in the residual quantiles. Most interestingly though, the reconstructed quantile distributions frequently display consecutive shifts between the sub-periods. In turn, equivalent changes occur in the residual quantiles of the simulated data. Often, but not always, the reconstructed quantile distributions shift to more negative values. This results, at some grid-points, in the following behavior of the residuals of simulated quantiles: residuals change from being negatively biased to being positively biased with an interval of nearly negligible residuals and, thus, an interval for which reconstructed and simulated quantiles appear to be consistent. These shifts suggest that the mean state of the reconstruction changes between the three sub-periods, and that the simulated distributions do not follow but usually feature a rather constant climatology. Patterns of residual quantiles are generally comparable between the different simulations. However, ensemble members may feature distinct residuals especially in the distributional tails. Obviously, the sample for each sub-period is small since we use non-overlapping decadal means with the full-period consisting of 85 data points.

For the same sample, rank counts confirm probabilistically the result of a generally over-dispersive ensemble by showing predominantly dome-shapes (Fig. 2, right panels). Compensating discrepancies in different sub-periods can imply consistency over the full period, e.g. opposite biases cancel each other out and the rank counts are approximately uniform over the full period (see, e.g. grid-point 11E 3S). Inconsistencies

CPD
9, 3789–3824, 2013

originate from different discrepancies at different locations and at the same location for different sub-periods.

Thus, according to probabilistic and climatological considerations the ensemble appears to be often over-dispersive relative to the field reconstruction. The too-dispersive ensemble-character agrees with the findings of Bothe et al. (2013) for the COSMOS-Mill-ensemble (Jungclaus et al., 2010). Furthermore, since the distributional evaluation suggests changes over time in the relation between reconstruction and simulation ensemble, we can infer that reconstruction and simulations often do not represent the same climate trajectory. Neither the single-model-ensemble (COSMOS-Mill) nor the past1000 multi-model-ensemble reliably represents the climate evolution suggested by the reconstruction. However, this may be due to the uncertainties associated with the verification data.

3.2 Consistency of indices

Since the North Pacific and North Atlantic are of particular interest in assessments of low-frequency climate variability and the field evaluation indicates at best limited consistency there (compare Fig. 1), we next consider surface air temperature indices for both domains (see Sect. 2.2 for details). Accounting for uncertainty in the reconstructions, the full-period residual distributions of simulated Pacific (PDO) and Atlantic (AMO) time-series arise as to some extent over-dispersive (Fig. 3a and c) mainly due to an overestimation of the tails especially for negative anomalies. Residuals are nearly negligible for some simulations. The 90 % envelope for a block-bootstrap approach marginally includes the zero line of consistency for the PDO. Thus, the sampling uncertainties prevent rejecting consistency.

For both indices, the reconstructed distributions for sub-periods shift from mainly positive temperature anomalies towards negative anomalies (Fig. 3b and d). The associated residual quantiles resemble the temporal development of residual grid-point-data quantiles. A slight cold bias in the early sub-period with especially large deviations for the cold tail changes to a generally over-dispersive relation in the latest sub-periods.

CPD
9, 3789–3824, 2013

Distributional changes between the last two sub-periods are less prominent for PDO compared to AMO.

The full-period rank counts are significantly over-dispersive for both PDO and AMO (Fig. 3e and g) according to the goodness-of-fit test, indicating that the ensemble is not probabilistically consistent with the reconstruction for both indices. The bootstrapped intervals confirm the rejection of consistency although only marginally for the AMO.

For the sub-periods, the simulation ensembles are significantly biased for both indices in the early and significantly over-dispersive in the central sub-period (Fig. 3f and h). Bias and spread are significant for the PDO in the last sub-period, but only bias is significant for the AMO then (Fig. 3f and h). Especially prominent are the over-dispersion for the late-period PDO estimates and the bias for the late-period AMO.

Thus, the regional indices confirm the field assessment result of a simulation-ensemble that tends to be over-dispersive relative to the global field reconstruction. Again, the simulation data do not reproduce the notable changes in the reconstructed distributions.

3.3 Consistency of regional reconstructions

We consider additional regional area-averaged temperature reconstructions to evaluate whether the mixed result relative to the Mann et al. (2009) field reconstruction is representative. Already the prominent uncertainty of climate reconstructions requires such additional evaluations.

We show only results for annual central European (Dobrovolný et al., 2010) and annual Southwestern North America (Wahl and Smerdon, 2012) temperatures starting from 1500 CE. Other regional reconstructions were assessed as well but are not discussed in depth. Accounting for uncertainties in the reconstructions, residual quantile distributions indicate often full-period over-dispersion for the decadal Central European temperature data (Fig. 4a). On the other hand, the data for Southwestern North America is mainly consistent (Fig. 4b). Nevertheless deviations occur for some simulations but are not significant. These include an over- as well as an under-estimation of the cold tail and an over-estimation of the warm tail. The differences among simulations are more diverse for climatological residual quantiles relative to the Southwestern North America reconstruction compared to the results concerning the large-scale indices and the grid-point data. Climatological relations can differ remarkably for different regional reconstructions as exemplified by the Central European and Southwestern North American data.

Rank histograms (Fig. 4c and d) indicate that the ensemble is probabilistically consistent with the Southwestern North American reconstruction, but the χ^2 goodness-of-fit test leads us (only just) to reject uniformity for the European data rank counts at the considered one-sided 90 % level (Fig. 4c and d). Similarly, the bootstrapped intervals in Fig. 4a–d do only just result in rejecting consistency for the European data but they in principle confirm the consistency for the American data. The bootstrapped envelope also highlights the high sampling variability. We note that over-dispersive deviations are much smaller for the Central European data than for the large-scale indices or the grid-point data.

Thus, the evaluation indicates better consistency of the ensemble relative to the two semi-millennial regional annual reconstructions than for either large-scale indices or grid-point data during the full period. On the other hand, analyses on additional millennial-scale reconstructions indicate usually stronger climatologically and probabilistically over-dispersive relations with, again, notable variations in consistency over time (not shown). These regional area-average data-sets often differ more strongly in their variability from the simulation ensemble than the central European and annual Southwestern North America data.

We note that the ensemble shows negligible under-dispersion relative to the European reconstruction if we exclude the uncertainties (not shown), but it indicates slight over-dispersion under uncertainties (Fig. 4a). One could argue that, for an ideal ensemble, such rather weak opposite deviations indicate a consistent ensemble and only an over-estimation of the target uncertainties (Persson, 2011, Appendix B; see also Hargreaves et al., 2011).

CPD
9, 3789–3824, 2013

4 Comparison of simulations and reconstructions – sources of disagreement

Significant distributional inconsistencies possibly render moot the evaluation of the agreement between simulations and reconstructions. In the best case, the associated uncertainties blur the common signals. In the worse case, the estimates do not represent the same distribution. Nevertheless, the mutual (dis)agreement sheds light on the shortcomings of models and reconstructions.

4.1 Global temperature fields

The lack of consistency of the past1000-ensemble is slightly less prominent compared to the COSMOS-Mill ensemble for the global temperature field for the full and the sub-periods and for the full and single tests. However, Bothe et al. (2013) used interannually resolved (but decadally smoothed) data and here we use non-overlapping decadal averages. For both ensembles, deviations of the simulations from the reconstruction differ strongly between different sub-periods and even include opposite deviations. It seems that overall over-dispersion is less prominent for the decadally resolved multi-model past1000-ensemble than for the interannually resolved COSMOS-Mill ensemble. The deviations visualized in Fig. 2 and their similarity between individual simulations in the past1000-ensemble suggest that this is more due to the temporal resolution than due to the ensemble characteristics.

To identify sources of disagreement we first consider mapped correlation coefficients between simulations and reconstructions (Fig. 5). Again we employ non-overlapping decadal means. Each simulated and reconstructed time-series represents one realization of a climate response to the employed radiative forcing perturbations modulated by the internal variability. We also expect differences in parametrisations and methodologies to affect the outcome. We consider correlation analysis as an example for common tools in studies comparing simulations and reconstructions. Finding significant correlations between individual simulations and the reconstruction indicates that both

CPD
9, 3789–3824, 2013

data-sets to some extent feature a similar signal but it does not give information about the origin of the signal or whether the origin is common in both data sets.

Indeed, mapped correlation coefficients suggest various degrees of agreement between individual simulations and the reconstruction (Fig. 5). Correlations are significant nearly everywhere for the ensemble mean (two-sided 99 % level) and CCSM4 (two-sided 90 % level) but less widespread for the other simulations. Most simulations correlate significantly negative with the reconstruction at some grid-points over Antarctica. All simulations correlate significantly over the western tropical Pacific, the subtropical North Pacific and the South Atlantic. The simulations and the global reconstruction do not agree on the, possibly externally forced, phasing of variations in Antarctica and the eastern and central tropical Pacific. We note that Mann et al. (2009) report a pronounced cold anomaly in the tropical Pacific for the Medieval Warm Period (MWP). Prominent gaps in significance are also visible for the ensemble mean and for CCSM4 over the sub-polar North Atlantic, the tropical Pacific, the Indian Ocean and central Eurasia. Similarities in the correlation-patterns may be interpreted as reflecting not only the intra-ensemble forcing-variability/-similarity but also the association between the models (compare Masson and Knutti, 2011).

Latitude-time plots of zonal means allow further comparison of the different data sets (Fig. 6). The reconstruction represents a near-global transition from positive anomalies in the first half (the MWP) to negative anomalies in the second half (Little Ice Age, LIA) of the considered 850 yr period (Fig. 6). The zonal means are possibly not representative in high southern latitudes due to data sparseness. The strongest warmth occurs at the beginning of the millennium. Episodic warmth interrupts the LIA during the 15th and 18th centuries and is generally confined south of 50° N.

The simulations neither capture the timing of the strongest warmth nor the near-global MWP-LIA transition. The ensemble generally displays near-stationary warm conditions. Short cold episodes related to assumed volcanic eruptions interrupt this warmth. Their timing, amplitude and spatio-temporal extent are similar in individual simulations. Weaker cold excursions reflect to some extent the variety of the employed

CPD
9, 3789–3824, 2013

forcings for reconstructed volcanic eruption properties (compare Schmidt et al., 2011). The ensemble mean differs most notably from the reconstruction in the lack of persistent northern hemispheric cold anomalies after about 1450 and in a stronger simulated cold signal in the 13th century. Otherwise it visually agrees well with the reconstruction.

We note that ensemble-mean correlation coefficients are often especially high (Fig. 5h) close to the proxy-locations employed by Mann et al. (2009). This implies stronger commonalities at those locations where our proxy-information about past climates are collected. That is, the similarity may allow inferring that simulations, reconstructions and underlying proxies as well as the forcing series relate to a similar underlying climate signal. Such inference is in accordance with the results of Schurer et al. (2013), Fernández-Donado et al. (2013) and Hind et al. (2012). On the other hand, the hypothesised common signal is concealed by the internal variability of the simulated climates and the additional sources of noise associated with simulations and reconstructions. We find no identifiable relation between the reconstruction and the simulations at the grid-point level.

4.2 Atlantic and Pacific indices

For the indices considered in the present study, the ensemble mean and the reconstruction evolve, by eye, similarly for the Atlantic and Pacific indices since about 1650 (Fig. 3i and j). Amplitudes agree less than tendencies. The most prominent example of differences in long-term trends leading to biased estimates is the different timing of medieval warmth. Note further the strong disagreement due to, on average, colder reconstructed indices from the 14th to 17th centuries for the PDO and from the 16th to 18th centuries for the AMO.

The indices display some intra-ensemble and ensemble-reconstruction agreement. Again we discuss correlations as example for common practices. Ensemble-mean indices correlate at $r \approx 0.5$ with the reconstructed ones. Correlations with the reconstructed index are larger than 0.5 for the PDO in FGOALS and for the AMO in CSIRO. Correlations among simulations larger than 0.5 are only found for the AMO and most

CPD
9, 3789–3824, 2013

prominently for MPI-ESM and CSIRO. PDO and AMO correlate strongest in the reconstruction, CSIRO and the ensemble mean ($r > 0.8$). If the analysis is repeated for globally detrended data, no strong correlations are seen between the reconstructed and simulated indices.

We did not discuss regional average indices in Bothe et al. (2013), but in both regions the COSMOS-Mill simulations displayed more variability than the reconstruction and, for the North Atlantic, the ensemble-consistency changed strongly between the considered sub-periods.

4.3 Regional temperatures

Figure 4 clearly displays that there is no common signal in the regional average time-series for Central Europe for individual simulations and reconstructions. This was similarly seen for the annually-resolved Central European temperature indices of the COSMOS-Mill ensemble. Obviously internal variability and methodological uncertainties dominate over the forced variability on the decadal and the inter-annual time scale for both ensembles. However, the COSMOS-Mill ensemble is consistent with the annual data of Dobrovolný et al. (2010) on the interannual time scale. Compared to the European data, the past1000-simulated Southwestern North American temperature series agree slightly better with the respective reconstruction for the non-overlapping decadal means. Considering the full ensemble, no common forced signal can be found. Thus we do not further comment on the accuracy of both datasets.

4.4 Further discussion

For the field and the index data, inconsistencies over the full period are due to a generally warmer start of the millennium in the reconstruction (compare Fig. 5). This would be mitigated for the analysis of ensemble-consistency and for the index-agreement between GISS and the reconstruction if the drift was not corrected for GISS-R (compare Schmidt et al., 2012). Decadal temperatures and their variability are more comparable in the period of the early LIA. Shifts in reconstructed quantiles towards more negative anomalies and in simulated residuals towards a more positive bias reflect the more pronounced reconstructed MWP-LIA transition and, thus, the differences in the long-term trends. Note that specific results are sensitive to the choice of the reference period. If we align the data sets to a different common period, specific relations are going to change relative to most reconstructions further highlighting the large discrepancies between the simulations and the reconstructions. The differences in estimates of cold anomaly quantiles reflect that, on the one hand, reconstructions possibly underestimate the cooling subsequent to large volcanic eruptions (Mann et al., 2012) while, on the other hand, models may be too sensitive to the subsequent radiative forcing anomaly (e.g. Anchukaitis et al., 2012).

In view of a possible impact of the choice of forcing inputs on the simulated data, one might think about partitioning the ensemble relative to the various combinations of forcings (Table 1). Only discriminating by the volcanic forcing, this results in two sub-ensembles including, respectively BCC, IPSL, CCSM4 and GISS-R25 (using the data by Gao et al., 2008) and MPI-ESM, CSIRO and GISS-R24 (using the Crowley data, see, e.g. Crowley and Unterman, 2012). Here we exclude the FGOALS data as it does not easily fit into these two categories but considers forcings as presented by Jones and Mann (2004) which are not explicitly included in the PMIP3-protocol (see Table 1 and Schmidt et al., 2011). Comparing the two ensembles, we see that the effect of the different volcanic forcing data sets is comparable to the effect of different model architectures inferred from the within sub-ensemble variations (not shown). This evaluation further implicates that the implementation strategy for the volcanic forcing data and the tuning of the model may influence the results as much as the choice of the forcing data. We note that Schmidt et al. (2013) report a radiative forcing twice as strong as expected from the Gao et al. (2008) data for the GISS-R simulations. They attribute this fact to the implementation of the volcanic forcing data in the model. We use the ensemble member GISS-R25 which employs the Gao et al. (2008) data. The

CPD
9, 3789–3824, 2013

simulated impacts of strong volcanic eruptions on temperature are among the largest but not generally exceptional (not shown, compare Fig. 6).

A number of possible additional and confounding factors may influence the proxies used to reconstruct the forcing data and the temperature data (e.g. precipitation, cloudiness, general circulation). Furthermore, the simulations possibly do not fully capture the influence of, e.g. the solar forcing due to deficiencies in the representation of atmospheric chemistry and to an only partially-resolved stratosphere. These issues become especially prominent prior to 1400 (e.g. Schurer et al., 2013). Correlations may also be dominated by short-lived episodes of large forcing which are commonly featured by simulations and reconstructions (compare Schurer et al., 2013). However, Bothe et al. (2013) also showed that it is inadequate to expect larger agreement and consistency closer to the present. Since the zonal means suggest similarities by filtering out regional differences, one might hypothesize that the lack of common signals between reconstructions and simulations at the grid-point level is solely due to the internal and local variability masking it.

5 Summary and conclusions

The CMIP5/PMIP3-past1000-ensemble is not generally consistent with the global temperature reconstructions by Mann et al. (2009) on a decadal time-scale. This holds for the probabilistic and the climatological assessments. Inconsistencies between reconstructions and simulations prevent reconciling both paleoclimate estimates. Our assessment of consistency over the last millennium can be biased towards being too optimistic if existing discrepancies between different multi-centennial sub-periods counterbalance each other.

The simulations and the reconstruction agree least in the tropical Pacific and the subpolar gyre region of the North Atlantic according to our evaluation, while agreement is largest in the sub-tropical Pacific and the South Atlantic. The large-scale significant

CPD
9, 3789–3824, 2013

correlations for some individual simulations and the ensemble mean indicate that the reconstruction and the simulation ensemble possibly include a common signal.

Robust conclusions require considering more than one data set and more than one parameter due to the large uncertainties. To this regard, the ensemble is also frequently over-dispersive relative to independent area-averaged regional temperature reconstructions. However, the ensemble is probabilistically consistent with the reconstructed annual temperatures for the Southwestern North America (Wahl and Smerdon, 2012).

The PMIP3-past1000 multi-model ensemble and the COSMOS-Mill single-model ensemble (Bothe et al., 2013) give very similar results with respect to their consistency, although differences exist for the diagnosed climatological deviations, which we attribute to different handling of volcanic forcing data. So, multi-model and single-model ensembles similarly lack consistency with the reconstructions. Thus, the uncertainty due to structural differences and parametrisations in the models does apparently not exceed the uncertainties associated with different forcing and initial conditions.

The PMIP3-past1000 simulation-ensemble and a selection of global and regional validation reconstruction targets are often not exchangeable climatologically and probabilistically. Therefore they should not be regarded as representing the same climate, i.e. they should not be compared under that implicit assumption.

These results imply the following:

1. The ensemble may be consistent with the verification data for either the full or for sub-periods at the grid-point level and for area-averaged data, but only few data show consistency on both time scales.

2. If consistency is diagnosed only for the full period, the ensemble and the reconstruction display a comparable amount of variability and a comparable climatological range over this period, but the long-term trends differ notably. We can also conclude that the variability differs between frequency bands, e.g. the reconstruction displays larger multi-centennial but smaller decadal variability and vice versa.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

CPD
9, 3789–3824, 2013

CMIP5/PMIP3-past1000 consistency

O. Bothe et al.

Furthermore, analyses depending on the background climate are hampered by the lack of sub-period-consistency.

3. If, on the other hand, the data is consistent for a sub-period, analyses on the dynamics may be valid over this period, but not necessarily in other periods. Even then we have to be careful since considering a different reference period climatology may lead to different results and the background climate may influence our assessment of the dynamics.

The deviations from consistency disclose the necessity of improvements for simulated estimates, their reconstructed forcing data and for climate-reconstructions. The large uncertainties render difficult any firm conclusions on past climate forcing and past climate variability (e.g. Hind et al., 2012).

Summarizing, probabilistic and climatological evaluations indicate consistency of the PMIP3-past1000 simulation-ensemble with the reconstructions in some regions and over certain sub-periods, but strong biases and/or dispersion-deviations arise in other regions and periods. This dominant feature of the analysis prevents reconciling the simulated and reconstructed time-series either with respect to a common naturally forced climate signal or with respect to an estimate of internal variability.

Supplementary material related to this article is available online at: http://www.clim-past-discuss.net/9/3789/2013/cpd-9-3789-2013-supplement.pdf.

Acknowledgements. We acknowledge comments by Eduardo Zorita and Jochem Marotzke, which lead to notable improvements on an earlier version of the manuscript as did two anonymous reviews. O. Bothe received funding by the DFG through the Cluster of Excellence CliSAP, University of Hamburg. D. Zanchettin was supported by the BMBF (research program "MiKlip", FKZ:01LP1158A). This work contributes to the MPI-M Integrated Project Millennium and CliSAP. We acknowledge the World Climate Research Programme's Working Group on

Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. This work benefited from the efforts of the Paleoclimatology Program at NOAA and its archive of reconstructions of past climatic conditions and forcings.

The service charges for this open access publication have been covered by the Max Planck Society.

References

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., Korner, 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. 3808

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996. 3793, 3794, 3795, 3798

Annan, J. and Hargreaves, J.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, doi:10.1029/2009GL041994, 2010. 3791, 3793

Annan, J., Hargreaves, J., and Tachiiri, K.: On the observational assessment of climate model performance, Geophys. Res. Lett., 38, L24702, doi:10.1029/2011GL049812, 2011. 3796

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, 2013. 3791, 3792, 3793, 3794, 3796, 3798, 3801, 3804, 3807, 3809, 3810

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 palaeoclimatic data, Nat. Clim. Change, 2, 417–424, doi:10.1038/nclimate1456, 2012. 3790

CPD
9, 3789–3824, 2013
CMIP5/PMIP3-past1000 consistency
O. Bothe et al.

Bradley, R. S.: High-resolution paleoclimatology, in: Dendroclimatology, Springer, 3–15, doi:10.1007/978-1-4020-5725-0_2, 2011. 3795

Brohan, P., Kennedy, J. J., Harris, I. Tett, S. F. B., and Jones, P. D.: Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850, J. Geophys. Res.-Atmos., 111, D12106, doi:10.1029/2005JD006548, 2006. 3798

Crowley, T. J.: Causes of climate change over the past 1000 years, Science, 289, 270–277, 2000. 3818

Crowley, T. J. and Unterman, M. B.: Technical details concerning development of a 1200-yr proxy index for global volcanism, Earth Syst. Sci. Data Discuss., 5, 1–28, doi:10.5194/essdd-5-1-2012, 2012. 3808

Crowley, T. J., Zielinski, G., Vinther, B., Udisti, R., Kreutz, K., Cole-Dai, J., and Castellano, E.: Volcanism and the little ice age, Pages News, 16, 22–23, 2008. 3818

Dima, M. and Lohmann, G.: A hemispheric mechanism for the Atlantic Multidecadal Oscillation, J. Climate, 20, 2706–2719, doi:10.1175/JCLI4174.1, 2007. 3797

Dobrovolný, P., Moberg, A., Brázdil, R., Pfister, C., Glaser, R., Wilson, R., Engelen, A., Limanówka, D., Kiss, A., Halíčková, M., Macková, J., Riemann, D., Luterbacher, J., and Böhm, R.: Monthly, seasonal and annual temperature reconstructions for Central Europe derived from documentary evidence and instrumental records since AD 1500, Climatic Change, 101, 69–107, doi:10.1007/s10584-009-9724-x, 2010. 3791, 3792, 3796, 3802, 3807, 3822

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. 3806

Gao, C., Robock, A., and Ammann, C.: Volcanic forcing of climate over the past 1500 years: an improved ice core-based index for climate models, J. Geophys. Res.-Atmos., 113, D23111, doi:10.1029/2008JD010239, 2008. 3808, 3818

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Mueller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850

to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5, J. Adv. Model. Earth Syst., doi:10.1002/jame.20038, in press, 2013. 3797
Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001. 3794
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. 3791, 3793, 3796, 3803
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. 3791
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. 3806, 3811
Hourdin, F., Foujols, M.-A., Codron, F., Guemas, V., Dufresne, J.-L., Bony, S., Denvil, S., Guez, L., Lott, F., Ghattas, J., Braconnot, P., Marti, O., Meurdesoif, Y., and Bopp, L.: Impact of the LMDZ atmospheric grid configuration on the climate and sensitivity of the IPSL-CM5A coupled model, Clim. Dynam., 40, 2167–2192, doi:10.1007/s00382-012-1411-3, 2012. 3797
Johnson, C. and Bowler, N.: On the reliability and calibration of ensemble forecasts, Mon. Weather Rev., 137, 1717–1720, doi:10.1175/2009MWR2715.1, 2009. 3793
Jolliffe, I. T. and Primo, C.: Evaluating rank histograms using decompositions of the chi-square test statistic, Mon. Weather Rev., 136, 2133–2139, doi:10.1175/2007MWR2219.1, 2008. 3793, 3794
Jones, P. and Mann, M.: Climate over past millennia, Rev. Geophys., 42, RG2002, doi:10.1029/2003RG000143, 2004. 3808, 3818
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. 3792, 3801
Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., Mikolajewicz, U., Notz, D., and von Storch, J.: Characteristics of the ocean simulations in

CPD
9, 3789–3824, 2013
CMIP5/PMIP3-past1000 consistency
O. Bothe et al.

MPIOM, the ocean component of the MPI-Earth system model, J. Adv. Model. Earth Syst., doi:10.1002/jame.20023, in press, 2013. 3797

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. 3796

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., 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, doi:10.1126/science.1177303, 2009. 3791, 3792, 3796, 3797, 3798, 3799, 3802, 3805, 3806, 3809

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, doi:10.1038/ngeo1394, 2012. 3808

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, doi:10.1175/2010MWR3129.1, 2011. 3791, 3793, 3794

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, L08703, doi:10.1029/2011GL046864, 2011. 3805

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikoajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the climate of a global model, J. Adv. Model. Earth Syst., 4, M00A01, doi:10.1029/2012MS000154, 2012. 3792

Meehl, G., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J., Stouffer, R., and Taylor, K.: The WCRP CMIP3 multi-model dataset: a new era in climate change research, B. Am. Meteorol. Soc., 88, 1383–1394, doi:10.1175/BAMS-88-9-1383, 2007. 3792

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973. 3794

Persson, A.: User Guide to ECMWF forecast products, Tech. rep., ECMWF, Reading, UK, 2011. 3803

Phipps, S. J., Rotstayn, L. D., Gordon, H. B., Roberts, J. L., Hirst, A. C., and Budd, W. F.: The CSIRO Mk3L climate system model version 1.0 – Part 2: Response to external forcings, Geosci. Model Dev., 5, 649–682, doi:10.5194/gmd-5-649-2012, 2012. 3797

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A., and Taylor, K. E.: Climate models and their evaluation, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., chap. 8, Cambridge University Press, 2007. 3795

Rougier, J. C., Goldstein, M., and House, L.: Second-order exchangeability analysis for multimodel ensembles, submitted, 2012. 3793

Sanderson, B. M. and Knutti, R.: On the interpretation of constrained climate model ensembles, Geophys. Res. Lett., 39, L16708, doi:10.1029/2012GL052665, 2012. 3792

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. 3790, 3795, 3799, 3806, 3808

Schmidt, G. A., LeGrande, A., and Healy, R.: "Interpretation of CMIP5 model ensemble", Interactive comment on "Constraining the temperature history of the past millennium using early instrumental observations" by P. Brohan et al., Clim. Past Discuss., C393–C398, 2012. 3797, 3807

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 paleo-climate comparisons to constrain future projections in CMIP5, Clim. Past Discuss., 9, 775–835, doi:10.5194/cpd-9-775-2013, 2013. 3791, 3792, 3795, 3799, 3808

Schurer, A., Hegerl, G., Mann, M. E., Tett, S. F., and Phipps, S. J.: Separating forced from chaotic climate variability over the past millennium, J. Climate, doi:10.1175/JCLI-D-12-00826.1, in press, 2013. 3806, 3809

Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, Geophys. Res. Lett., 36, L19704, doi:10.1029/2009GL040142, 2009. 3818

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. 3791

CPD
9, 3789–3824, 2013

CMIP5/PMIP3-past1000 consistency

O. Bothe et al.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, doi:10.1029/2000JD900719, 2001. 3791, 3793
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. 3790, 3792
Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philos. T. Roy. Soc. A, 365, 2053–2075, doi:10.1098/rsta.2007.2076, 2007. 3792
Tsonis, A. A., Wang, G., Swanson, K. L., Rodrigues, F. A., and da Fontura, C. L.: Community structure and dynamics in climate networks, Clim. Dynam., 37, 933–940, 2011. 3797
Vieira, L. E. A., Solanki, S. K., Krivova, N. A., and Usoskin, I.: Evolution of the solar irradiance during the Holocene, Astr. Astrophys., 531, A6, doi:10.1051/0004-6361/201015843, 2011. 3818
Wahl, E. R. and Smerdon, J. E.: Comparative performance of paleoclimate field and index reconstructions derived from climate proxies and noise-only predictors, Geophys. Res. Lett., 39, L06703, doi:10.1029/2012GL051086, 2012. 3791, 3796, 3802, 3810, 3822
Wilks, D.: On the reliability of the rank histogram, Mon. Weather Rev., 139, 311–316, doi:10.1175/2010MWR3446.1, 2011. 3793
Wilson, R., D'Arrigo, R., Buckley, B., Büntgen, U., Esper, J., Frank, D., Luckman, B., Payette, S., Vose, R., and Youngblut, D.: A matter of divergence: tracking recent warming at hemispheric scales using tree ring data, J. Geophys. Res.-Atmos., 112, D17103, doi:10.1029/2006JD008318, 2007. 3795
Zanchettin, D., Rubino, A., Matei, D., Bothe, O., and Jungclaus, J. H.: Multidecadal-to-centennial SST variability in the MPI-ESM simulation ensemble for the last millennium, Cilm. Dynam., 5–6, 1301–1318, doi:10.1007/s00382-012-1361-9, 2013. 3798
Zhou, T., Li, B., Man, W., Zhang, L., and Zhang, J.: A comparison of the Medieval Warm Period, Little Ice Age and 20th century warming simulated by the FGOALS climate system model, Chinese Sci. Bull., 56, 3028–3041, doi:10.1007/s11434-011-4641-6, 2011. 3796

CPD
9, 3789–3824, 2013

CMIP5/PMIP3-past1000 consistency

O. Bothe et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◂ | ▸
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Table 1. Selected climate model simulations and their acronyms, the institutes of origin and the respective solar and volcanic forcing data sets. Full references are (from top to bottom): Vieira et al. (2011), Gao et al. (2008), Steinhilber et al. (2009), Crowley (2000), Jones and Mann (2004) and Crowley et al. (2008).

Model (Acronym)	Institute	Solar	Volcanic
bcc-csm1-1 (BCC)	Beijing Climate Center, China Meteorological Administration	Vieira	Gao
CCSM4	National Center for Atmospheric Research	Vieira	Gao
CSIRO-Mk-3L-1-2 (CSIRO)	University of New South Wales	Steinhilber	Crowley (2008)
FGOALS-gl (FGOALS)	State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences	Crowley (2000), Jones and Mann (2004)	Crowley (2000), Jones and Mann (2004)
GISS-E2-R (GISS-R24)	National Aeronautic and Space Administration, Goddard Institute for Space Studies	Vieira	Crowley (2008)
GISS-E2-R (GISS-R25)	National Aeronautic and Space Administration, Goddard Institute for Space Studies	Vieira	Gao
MPI-ESM-P (MPI-ESM)	Max Planck Institute for Meteorology	Vieira	Crowley (2008)
IPSL-CM5A-LR (IPSL)	Institut Pierre Simon Laplace des sciences de l'environnement	Vieira	Gao

CPD
9, 3789–3824, 2013
CMIP5/PMIP3-past1000 consistency
O. Bothe et al.

CPD
9, 3789–3824, 2013

CMIP5/PMIP3-past1000 consistency

O. Bothe et al.

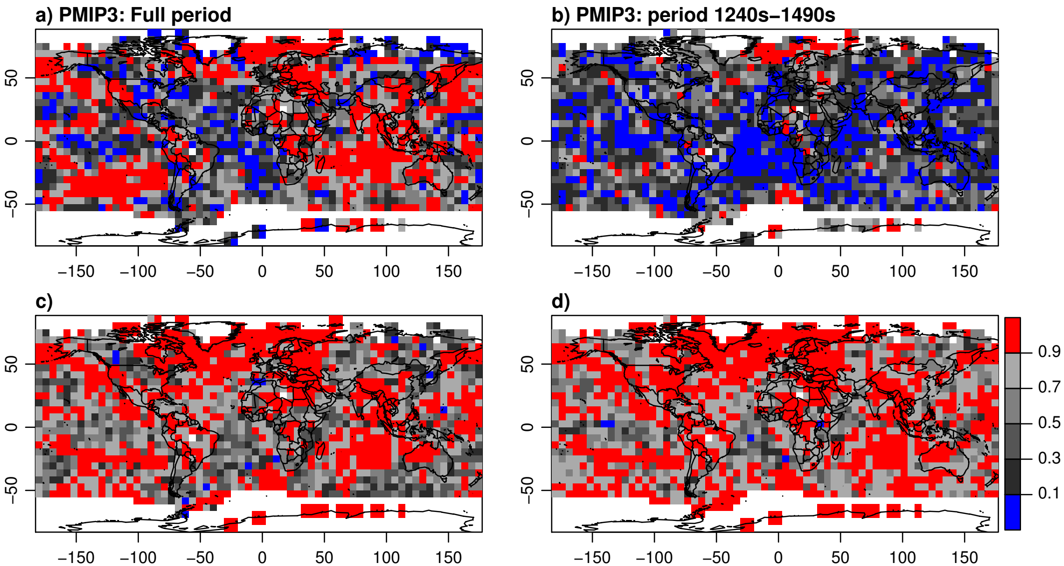

Fig. 1. Global assessment of the goodness-of-fit test for the field data considering uncertainties in the verification target. Plotted are lower p values. In the upper row: full χ^2 test, in the lower row: maximum of p values for single deviation tests for bias and spread. Blue smaller than 0.1, dark to light gray in steps of 0.2 within the range between 0.1 and 0.9, red larger than 0.9. Red means rejection of the uniform null hypothesis. **(a, c)** full period, **(b, d)** for the decades from the 1240s to the 1490s.

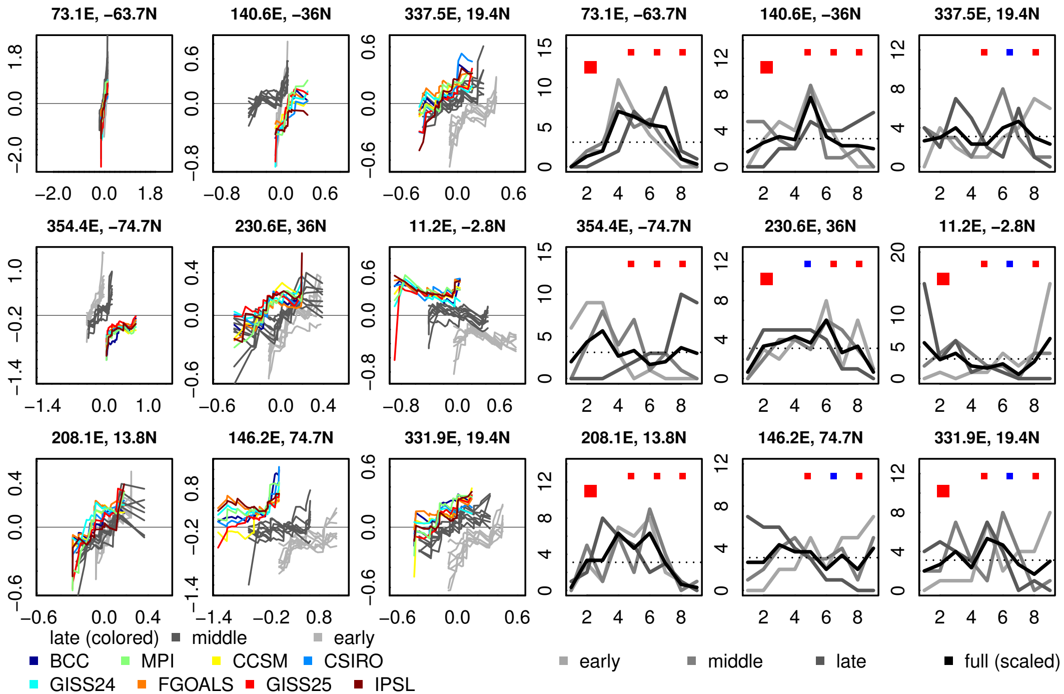

Fig. 2. Grid-point analysis of ensemble consistency for three sub-periods: 1000s–1270s, 1280s–1550s, 1560s–1830s). Left three columns: residual quantile-quantile plots for a selection of grid-points for the first (light gray), second (dark gray) and third (colored) sub-periods. Right three columns: rank histogram counts for the selection of grid-points for the three sub-periods (first to last, light to dark gray) and the full period (black, scaled to match frequencies in sub-periods). Large (small) red squares mark grid-points where spread or bias deviations are significant over the full (from left to right the first to third sub-)period. Blue squares mark deviations which are not significant.

CPD
9, 3789–3824, 2013

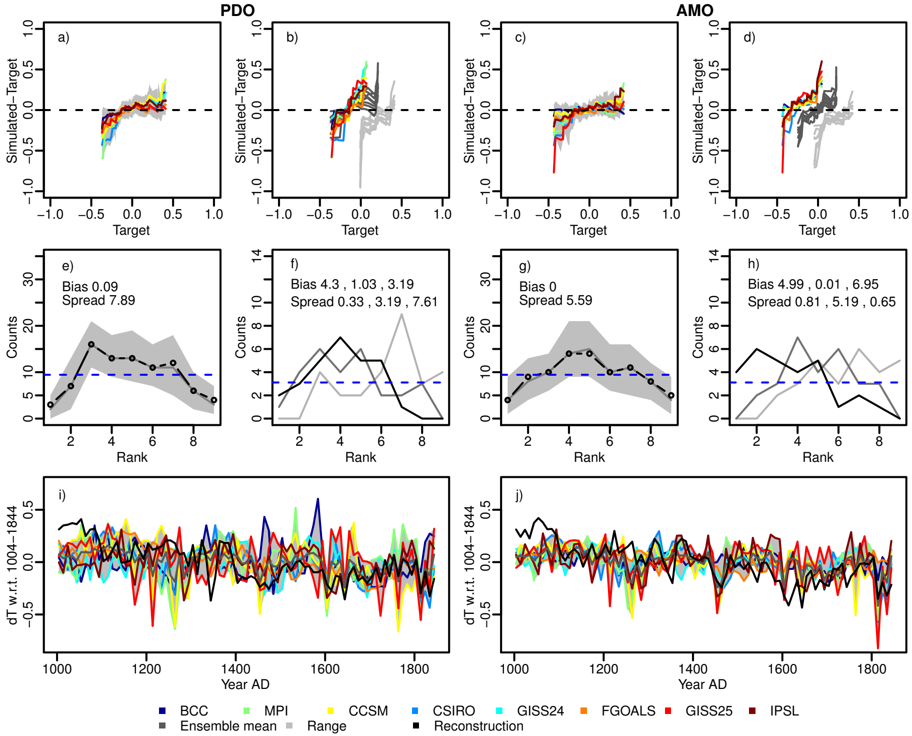

Fig. 3. Consistency of the indices for the North Pacific (PDO) and North Atlantic (AMO) regions. **(a–d)** Residual quantile-quantile plots for **(a, c)** the full period and **(b, d)** three sub-periods (defined as for Fig. 2) of 28 records (early, light gray, middle, dark gray, late, colored). **(e–h)** Rank histogram counts for **(e, g)** the full period and **(f, h)** the three sub-periods (light gray to black). Numbers are the χ^2 statistics for the periods. In **(f, h)** numbers refer, from left to right, to the early to late sub-periods. Blue horizontal lines give the expected average count for a uniform histogram. **(i, j)** Time series of the indices constructed from non-overlapping decadal means. Color-code as in legend except for shading. Shading for residual-quantiles and rank-counts **(a, c, e, g)** gives the 90 % envelope of block-bootstrapping 2000 replicates of block-length 5.

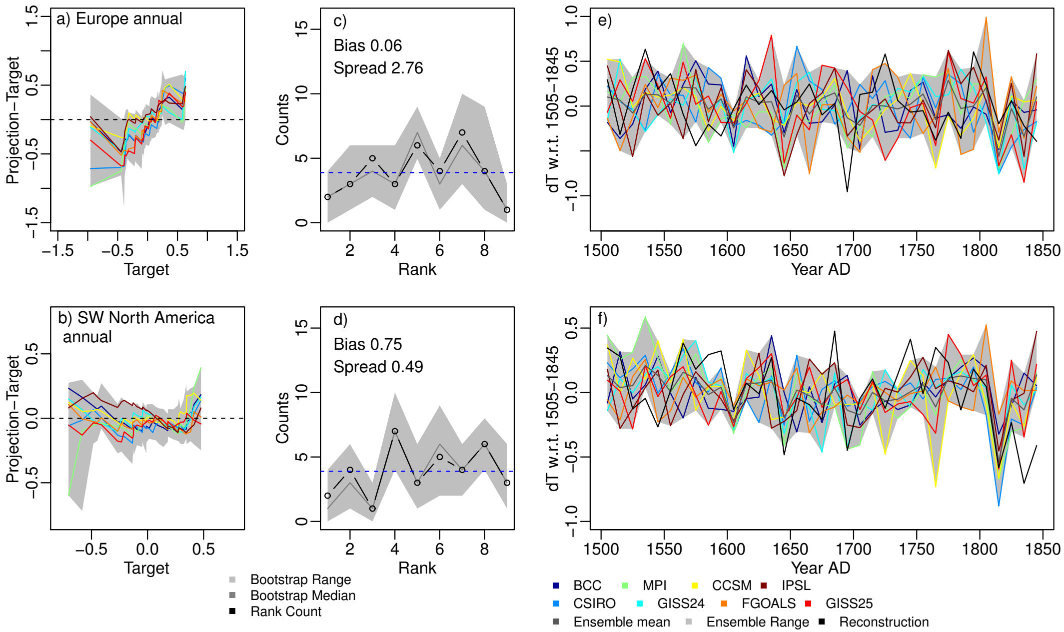

Fig. 4. Full-period residual quantile-quantile plots (left panels), rank counts (middle panels) and time series plots (right panels) for the reconstructions by (top panels) Dobrovolný et al. (2010) of Central European annual temperature and (bottom panels) Wahl and Smerdon (2012) of Southwestern North America annual temperature. For details on the representation see the caption of Fig. 3.

CPD
9, 3789–3824, 2013

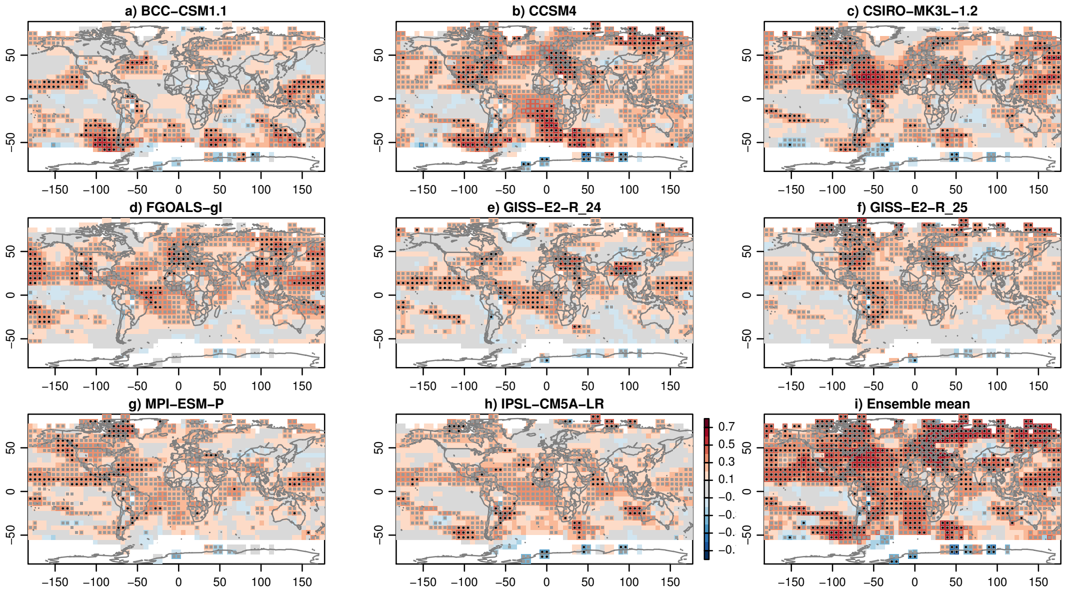

Fig. 5. Mapped grid-point correlation coefficients between surface air temperature series from the considered simulations and from the reconstruction. See panel titles for individual simulations. Ensemble mean in **(i)**. Gray (black) dots mark two-sided 90 % (99 %) confidence.

CPD
9, 3789–3824, 2013

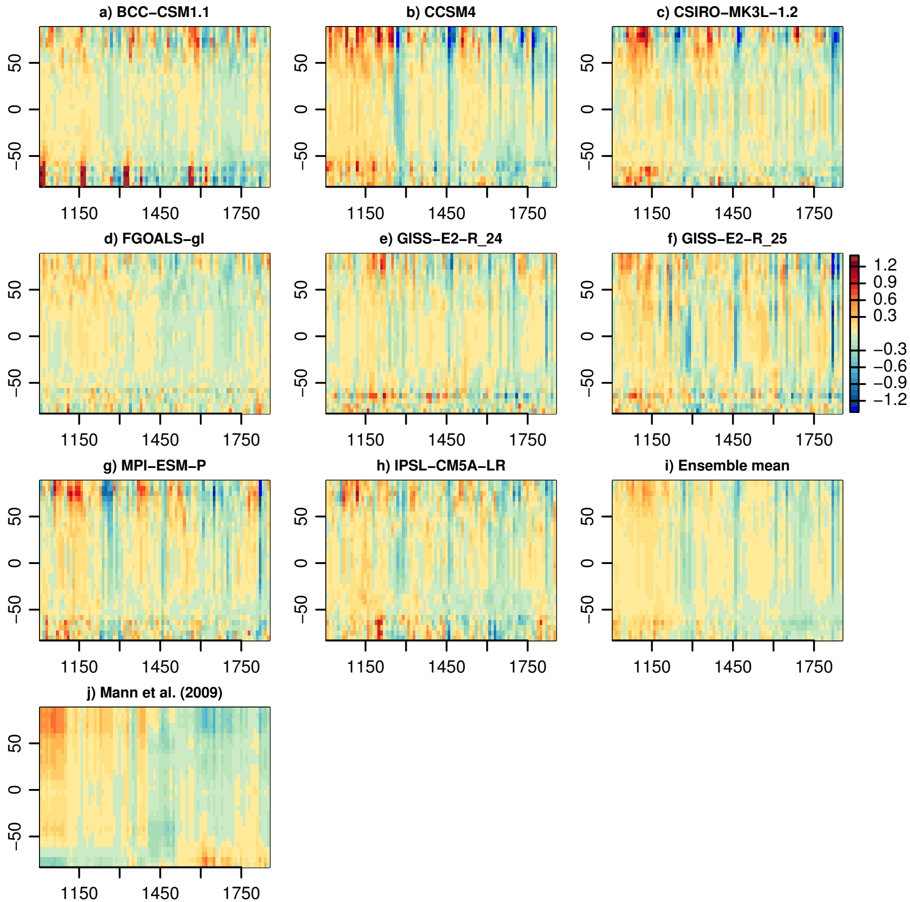

Fig. 6. Time-latitude-plots of full-field zonal mean temperature anomalies with reference to the analysis period. See panel titles for individual simulations. Ensemble mean in **(i)** and reconstruction in **(j)**.

CPD
9, 3789–3824, 2013

CMIP5/PMIP3-past1000 consistency

O. Bothe et al.