Clim. Past Discuss., 8, 2409–2444, 2012
www.clim-past-discuss.net/8/2409/2012/
doi:10.5194/cpd-8-2409-2012

Climate
of the Past
Discussions

Climate of the last millennium: ensemble consistency of simulations and reconstructions

O. Bothe[1,2], J. H. Jungclaus[1], D. Zanchettin[1], and E. Zorita[3,4]

[1]Max-Planck-Institute for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany
[2]University of Hamburg, KlimaCampus Hamburg, Germany
[3]Institute for Coastal Research, Helmholtz Centre Geesthacht, Germany
[4]Bert Bolin Centre for Climate Research, University of Stockholm, Sweden

Received: 31 May 2012 – Accepted: 5 June 2012 – Published: 22 June 2012

Correspondence to: O. Bothe (oliver.bothe@zmaw.de)

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

Abstract

Are simulations and reconstructions of past climate and its variability comparable with each other? We assess if simulations and reconstructions are consistent under the paradigm of a statistically indistinguishable ensemble. Ensemble consistency is assessed for Northern Hemisphere mean temperature, Central European mean temperature and for global temperature fields for the climate of the last millennium. Reconstructions available for these regions are evaluated against the simulation data from the community simulations of the climate of the last millennium performed at the Max Planck Institute for Meteorology.

The distributions of ensemble simulated temperatures are generally too wide at most locations and on most time-scales relative to the employed reconstructions. Similarly, an ensemble of reconstructions is too wide when evaluated against the simulation ensemble mean.

Probabilistic and climatological ensemble consistency is limited to sub-domains and sub-periods. Only the ensemble simulated and reconstructed annual Central European mean temperatures for the second half of the last millennium demonstrates consistency.

The lack of consistency found in our analyses implies that, on the basis of the studied data sets, no status of truth can be assumed for climate evolutions on the considered spatial and temporal scales and, thus, assessing the accuracy of reconstructions and simulations is so far of limited feasibility in pre-instrumental periods.

1 Introduction

Inferences about the spatio-temporal climate variability in periods without instrumental coverage rely on two tools: (i) reconstructions from (e.g.) biogeochemical and cultural (e.g. documentary) data that approximate the climate during the time of interest at a certain location in terms of a pseudo-observation; (ii) simulators (that is, models) of varying complexity that produce discretely resolved spatio-temporal climate variables considered to represent a climate aggregation over regional spatial scales. Confidence in the inference of a past climate state requires reconciling both estimates in terms of accuracy and reliability. In case of an ensemble of estimates, we have to evaluate the consistency of the ensemble with relevant validation data.

Similar to measurements by instrumental sensors, our pseudo-observations by proxies or paleo-sensors (as coined by Braconnot et al., 2012) are subject to "measurement" uncertainty. Uncertainties enter our reconstructions, among other ways, through the dating of the non-climate observation, the transfer function and the assumption of a relatively stable "proxy"-climate relationship through time (e.g. Wilson et al., 2007; Bradley, 2011). Simulated climate estimates are uncertain within the range of the mathematical and numerical approximations of physical and biogeochemical processes (Randall et al., 2007). Additional uncertainty comes from the reconstructions of the external factors driving the climate system simulation. These again are subject to dating and transfer uncertainty (Schmidt et al., 2011) resulting in diverse estimates of past solar (e.g. Steinhilber et al., 2009; Shapiro et al., 2011; Schrijver et al., 2011) and volcanic (e.g. Gao et al., 2008; Crowley and Unterman, 2012) variations.

If no status of "truth" can be assigned since, for example, we have no independent and reliable observational knowledge in the pre-instrumental period, the assessment of the statistical consistency provides an objective measure of confidence in our two tools. Thus, if we have an ensemble of simulations (reconstructions) we have to define a representation of the status of "truth" from the available reconstructions (simulations). For a specific task at hand, the analysis of consistency identifies whether the simulated and reconstructed climate estimates can be considered to be compatible realizations of an unknown "true" distribution, though not necessarily identical with it (Annan et al., 2011). Reconstructions and simulations are therefore treated as different but equitable hypotheses. Ensembles of hypotheses are available for northern hemispheric mean temperature reconstructions (Frank et al., 2010) and for the PMIP3-compliant Community Simulations of the last millennium (COSMOS-Mill, Jungclaus et al., 2010) allowing the assessment of the consistency of reconstructions and simulations within the framework of a statistically indistinguishable ensemble (Toth et al., 2003). Annan and Hargreaves (2010) and Hargreaves et al. (2011) discuss, respectively, the reliability of the CMIP3 ensemble and the ensemble consistency of the PMIP1/2 (Joussaume and Taylor, 2000; Braconnot et al., 2007) simulations in terms of this probabilistic interpretation. We adopt the Annan and Hargreaves (2010) approach to assess the mutual consistency among the ensembles of reconstructed and simulated estimates of northern hemispheric mean temperature for the last millennium. We further evaluate the consistency of temporal evolutions over the last millennium of the COSMOS-Mill ensemble with reconstructions for Central European mean temperature (Dobrovolný et al., 2010) and a temperature field reconstruction (Mann et al., 2009).

The following analysis is similar to the ensemble forecast verification in numerical weather prediction (Toth et al., 2003) and extends the application of the paradigm of statistical indistinguishability in the climate modelling context from climate means (Annan and Hargreaves, 2010; Hargreaves et al., 2011) to temporally varying climate trajectories. Probabilistic reconstruction-simulation consistency is assessed over the pre-industrial period of the last millennium using rank histograms (e.g. Anderson, 1996) and the decomposition of the χ^2-statistic (Jolliffe and Primo, 2008). The restrictions of the approach are considered by presenting residual quantile-quantile plots (Marzban et al., 2010; Wilks, 2010) to evaluate the climatological consistency. The methods are discussed in Sect. 2. Section 3 presents results concerning the consistency of reconstructions and simulations, and the sensitivity of the chosen approach is discussed in Sect. 4.

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

2 Methods and data

2.1 Methods

An ensemble of (climate) estimates can be validated either by considering individually the accuracy of each ensemble member against the "true" observation or by evaluating the reliability of the full ensemble, that is the compliance between "true" and ensemble estimated probability distributions (e.g. Marzban et al., 2010). Considering the multiple sources of uncertainties in paleo-climate reconstructions and simulations, assessing ensemble consistency objectifies our evaluation of ensemble accuracy. In the following, if we mention a "truth" or a "true" data set, this can only represent an uncertain approximation of the observable truth.

The reliability of a probabilistic ensemble is commonly validated under the paradigm of statistical indistinguishability by ranking true observational data against the ensemble data (Anderson, 1996; Jolliffe and Primo, 2008; Annan and Hargreaves, 2010; Marzban et al., 2010; Hargreaves et al., 2011). True and ensemble-simulated data are sorted by value and the calculated ranks counted and plotted as a rank histogram (Anderson, 1996).

A null hypothesis of a common overarching distribution for truth and ensemble implies equiprobable outcomes and the ranking should result in a uniform, flat histogram. For a "reliable" ensemble, observed and ensemble estimated (e.g. forecasted) frequencies agree (Murphy, 1973). Note, however, that a flat histogram of ranks does not necessarily imply reliability (see discussions by e.g. Hamill, 2001; Marzban et al., 2010).

Already visually, rank histograms assist in identifying discrepancies between the simulated probabilistic ensemble and the truth. If the truth is sampling from a distribution narrower (wider) than the ensemble, thus the spread of the ensemble is overly wide (too narrow), the rank histogram will appear dome-shaped (u-shaped). Too wide (narrow) ensembles are referred to as over-(under-)dispersive. If the ensemble is biased to positive (negative) values, a negative (positive) trend is seen in the rank counts. The "flatness" of the histogram can be assessed by a χ^2 goodness-of-fit test. Decomposing the test statistic enables tests for individual deviations from flatness; Jolliffe and Primo (2008) present a comprehensive delineation. In mapping spatial fields of verification ranks for climatological periods of interest (Sects. 3.2.1 and 3.2.2), individual low ranks of the truth hint to an overestimation of the climate parameter by the ensemble, whereas high ranks imply a negative bias in the simulation ensemble.

Meaningful statistics require to account for dependencies in the data (Jolliffe and Primo, 2008; Annan and Hargreaves, 2010) by e.g. evaluating the effective degrees of freedom in the time series. A higher number of degrees of freedom essentially leads to a higher chance of rejecting the hypothesis of uniformity. If ensemble and verification data are smoothed (as for the global data by Mann et al., 2009), either the sample size or the expected number of rank counts may be small compared with the theoretical requirements (but see e.g. Bradley et al., 1979, and references therein).

In assessing consistency for time series, temporal correlations in the data may further affect the structure of the rank histograms (Marzban et al., 2010; Wilks, 2010). Accounting for the sampling variability reduces the risk of drawing erroneous conclusions from the rank counts. We display, for area-averaged time series, quantile statistics of block-bootstrapped rank histograms (Marzban et al., 2010; Efron and Tibshirani, 1994). We apply a block length of 50 yr, calculate 2000 bootstrap replicates and display 0.5, 50 and 99.5 percentiles which also allows for a secondary test of uniformity.

The rank histogram approach further assumes that the true validation data includes an error (Anderson, 1996), which has to be included in the ensemble data. If the reconstructions are reported with an uncertainty estimate, this is used to inflate the simulated data.

Marzban et al. (2010, see also Wilks, 2010) recommend to evaluate the climatological component of reliability using residual quantile-quantile plots (r-q-q plots). Similar to common quantile-quantile plots, the estimated climatological quantiles are assessed against the true quantiles. Displaying the differences between the simulated distribution quantiles and the true quantiles emphasizes deviations in the climatological distributions. Biases result in a horizontal displacement from zero in the r-q-q plots, and climatological over- and under-dispersion (too wide or too narrow distributions) relate to positive or negative slopes (Marzban et al., 2010).

2.2 Data

We employ the ensemble of the COSMOS-Mill simulations for the last millennium performed with the Max Planck Institute Earth System Model (MPI-ESM) based on ECHAM5, MPI-OM, a land-surface module including vegetation (JSBACH), a module for ocean biogeochemistry (HAMOCC) and an interactive carbon cycle; details of the simulations have been published by Jungclaus et al. (2010). The set specifically includes single forcing simulations for volcanic, strong solar and weak solar forcing, five full-forcing simulations with weak solar forcing and three full-forcing simulations with strong solar forcings (full ensemble: eleven members). We include the single forcing simulations as valid hypotheses about the pre-industrial climate trajectory assuming that uncertainty is high in the respective forcing series and in our knowledge of the influence of the forcing components on the pre-instrumental climate. If a strong or weak ensemble is mentioned, this consists of the respective full-forcing simulations with strong and weak solar forcing.

Considered reconstructions are a regional annual temperature series for Central Europe (Dobrovolný et al., 2010), the ensemble data for annual Northern Hemisphere temperature by Frank et al. (2010) and the global temperature field reconstruction by Mann et al. (2009). For the Frank et al. (2010) data, we reverse the approach to study additionally the consistency of a reconstruction sub-ensemble with respect to the simulation ensemble mean; we use the sub-ensemble calibrated to the period 1920 to 1960. Spatial field data are interpolated on a 5 × 5 degree grid. As our interest is in the consistency of paleoclimate reconstructions and simulations for the last millennium, anomalies are taken with respect to the common period of reconstructions and simulations but excluding the period of overlap with the modern observations: (i) for the European temperature time series (period 1500 to 1854) with respect to the mean from 1500 to 1849, (ii) for the Northern Hemisphere temperature series for and with respect to the period 1000 to 1849, and (iii) for the decadal smooth global field the records for the years 805 to 1845 with respect to the mean for 800 to 1849.

2.3 Discussion of the chosen approach

The simulation-reconstruction-consistency can possibly be evaluated on three levels of resolution: area-averaged time series, gridded spatio(-temporal) data and individual grid points of the gridded data. Results may differ between these and it is not obvious at which level the consistency should be largest. Even if we find an ensemble of simulations to be consistent at the grid point level, we cannot say whether the covariance between individual grid points or within the whole field is consistent with the true covariability.

Uniformity in rank histograms may result from opposite biases or opposite deviations in spread in different periods or areas which cancel out (Hamill, 2001). On the other hand, indications of a too narrow ensemble may as well result from different biases in different periods. Temporal correlations in the data can result in premature rejection of flatness (Marzban et al., 2010). Using bootstrapped estimates or analysing different sub-periods at individual grid points helps to address these problems. We also follow Marzban et al. (2010) in displaying residual quantiles. Similar caveats exist for these climatological anomaly distributions.

Although the data sets are assumed to represent annually resolved values, this is not necessarily valid. If the target/truth is an ensemble mean, the target displays reduced inter-annual variability compared to the ensemble members. This has to be taken into account in interpreting the results. It is likely that using an ensemble mean as truth will change the ensemble consistency. Considering an error in the truth can compensate such problems. If reconstruction and simulation ensemble estimates are thought to include the same externally forced variability, the true ensemble mean should essentially recover the forced signal within the propagated uncertainties, and the probabilistic ensemble estimates (including the uncertainty of the truth) should reliably represent the true distribution. Similarly, members of the reconstruction ensemble are to some extent time-filtered and by construction exhibit reduced variability on inter-annual time-scales. As the properties differ for the reconstruction ensemble members, this filtering is not considered. On the other hand, the decadal smoothing of the global field data (Mann et al., 2009) is taken into account by using decadal moving means for the simulation ensemble data.

3 Results

We evaluate the ensemble consistency of the COSMOS-Mill simulation ensemble for area-averaged and grid point time series with respect to temperature reconstructions. In principle, all levels of spatial resolution are of interest, as the spatial and temporal availability of proxy records may hinder reconstructions on one of these levels and, yet, be sufficient for climate reconstructions on another. Implications and origins of found consistency or lack thereof are discussed.

3.1 Area-averaged time series

3.1.1 Ensemble consistency of area-averaged estimates

Figure 1 displays the data time series and their variability together with the range of the ensembles. Their probabilistic consistency is illustrated by Fig. 2 and the climatological component of consistency by Fig. 3. The bottom (top) rows of Figs. 2 and 3 do (do not) account for the error in the verification target.

No probabilistic differences arise between the ensemble simulated and reconstructed estimates for the Central European temperature (Fig. 2a), if the verification series is assumed to be perfect without error. Similarly, under such an assumption, the reconstruction sub-ensemble for the northern hemispheric mean temperature and the ensemble mean simulated Northern Hemisphere temperature are compatible (Fig. 2e). On the other hand, the simulation ensemble estimates for the Northern Hemisphere temperature are from a notably too wide probabilistic distribution relative to the ensemble mean reconstruction (Fig. 2c). The bootstrapped ranks (shading in Fig. 2) confirm this assessment. Although notable deviations may occur in the end ranks for the simulated European and reconstructed hemispheric temperature ensembles, they are not unlikely with respect to a uniform outcome.

Uncertainty estimates for the target data time series are the reported standard errors for the Central European temperature data (Dobrovolný et al., 2010) and the spread of the mutual ensembles for the Northern Hemisphere data. Accounting for these "errors" in the "verification" data alters the result for the reconstruction ensemble. The ranks in Fig. 2f clearly display strong over-dispersion, that is, the ensemble mean simulation populates too often the central ranks of the histogram. This behavior is also found for the ensemble mean reconstruction in Fig. 2d. The bootstrapped ranks and the goodness-of-fit test unambiguously indicate a lack of consistency due to over-dispersive distributions for the hemispheric data.

No large changes are found in the ranks for the European temperature data (Fig. 2b) and the 99 % range of the bootstrapped ranks is still compatible with a flat histogram. Contrarily, the presented χ^2 test gives significant p-values for spread-deviations, which highlights the problem of sampling variability and the strictness of the χ^2 test.

Similarly, the residual quantiles of the climatological distributions in Fig. 3a agree generally well for simulated and reconstructed European temperatures, although the simulations underestimate very warm annual anomalies and overestimate very cold ones. The time series in Fig. 1a relates the underestimation of the warm anomalies particularly to reconstructed extreme warmth in the mid 16th century. The overestimation of cold anomalies is more frequent but originates from only few ensemble members (Fig. 3a). If we include the error estimates, a slight slope occurs in the residual quantiles indicating that the simulations may sample from a slightly too wide distribution; the warmth in the 16th century remains exceptional.

Larger climatological deviations between the simulation ensemble and the reconstructions occur for the Northern Hemisphere temperature data (Fig. 3b, c). Independent of considerations on the reconstruction uncertainty, the simulation ensemble gives overly wide distributions. Similarly, the reconstruction ensemble overestimates the range of variability when compared to the simulation ensemble mean. While this again is in principle independent of the uncertainty in the truth, the deviations are largest in the positive anomaly quantiles if uncertainties are included.

Considering the two full-forcing simulation sub-ensembles separately (five simulations with weak, three with strong solar forcing) confirms the results with respect to the European temperature data although both ensembles display specific behaviors (not shown). If uncertainties in the truth are accounted for, the weak solar full-forcing ensemble is unambiguously consistent with the European reconstructions, whereas the strong solar forcing ensemble is slightly too wide. The spread is significant according to the goodness-of-fit test, but the bootstrapped ranks suggest that this may be due to sampling variability. The residual quantiles do not differ too much between both ensembles as seen in Fig. 3 (red, weak ensemble, blue, strong ensemble). Relative to the Northern Hemisphere temperature (not shown), both full-forcing sub-ensembles are significantly too wide according to the goodness-of-fit test, but the bootstrapped ranks generally include the possibility of a uniform histogram. The residual quantiles display strong deviations for the strong forcing ensemble (compare Fig. 3). Reversing the verification task and considering errors in the truth, the reconstruction ensemble distribution is too wide relative to the weak forcing ensemble but is consistent relative to the strong forcing ensemble (not shown).

The climatological assessment puts the probabilistic evaluation into perspective as it points to very strong deviations for the Northern Hemisphere mean temperatures. Bootstrapped residuals generally enclose the zero line for flatness, if the error in the truth is not considered, but deviations are outside the 99 % range for the positive tails otherwise. The reconstruction quantile residuals relative to the full simulation ensemble mean quantiles (Fig. 3e, f) present an amplified picture of the deviations relative to the two sub-ensembles.

Thus, verification of the simulation ensemble suggests that it is generally too wide compared to the employed area-average-reconstruction time series. Similarly, the reconstruction ensemble describes an over-dispersive distribution compared to the simulation ensemble mean. Strong discrepancies arise not only with respect to the probabilistic analysis but also in the climatological assessment. These, however, do not challenge the consistency of the Central European temperature estimates. On the other hand, the reconstruction ensemble displays strong deviations relative to the full and the single simulation ensemble means whereas the probabilistic assessment indicates consistency of the reconstruction ensemble relative to the strong solar forcing simulation ensemble mean. If 50 yr moving average series are considered for the hemispheric data, the general result remains that strong differences are seen probabilistically and/or climatologically between pairs of simulation ensemble and reconstruction.

3.1.2 Addressing origins of the lack of consistency

Figure 1 displays (i) that the European data for the simulations and the reconstruction cover a similar range and show similar variability, (ii) that the hemispheric reconstruction ensemble mean varies less than the simulation ensemble and displays different temporal evolution, as does (iii) the hemispheric simulation ensemble mean (compared to the reconstruction ensemble), which on the other hand is in the range of variability of the reconstruction ensemble. However, under the uncertainties associated with climate reconstructions, climate simulations and the forcing reconstructions, even such strongly differing estimates may be probabilistically and climatologically compatible with one another.

The scientific interest is to reconcile the simulated and reconstructed estimates of a climate close to the current, whose variations are only due to internal variability and natural, external forcings (Braconnot et al., 2012). The above analyses add estimates of the consistency of reconstructions and simulations, which can be viewed as measures of their comparability.

Thus, although the inset in Fig. 1 shows that European temperature evolves notably different before 1800 in the ensemble simulations and in the reconstruction, both datasets are in the above sense comparable. That is, the strong differences in the 18th century (or similarly the late 1500s) are likely compatible with our knowledge about internal and externally forced climate variability.

On the other hand, the distributions differ between the northern hemispheric temperature reconstruction ensemble mean and the full simulation ensemble, if we consider the uncertainty in the verification ensemble mean reconstruction. The time series clarify that part of the over-dispersive character of the ensemble may relate (i) to biases in the periods 1000 to 1300 and 1500 to 1650, where reconstructions and simulations evolve to some extent oppositely and to (ii) less warming in the reconstruction verification in the 18th century. The same biases act oppositely in the mutually reverse assessment and also influence the assessment of low frequent smoothed versions of the data. This is mostly, but not only, due to the evolution of the strong solar full-forcing simulation ensemble.

Figure 1 further shows that the considered ensembles of estimated temperature anomaly series generally enclose the verification data (Fig. 1a–c), but they often overestimate inter-annual variability (Fig. 1d–f). Verification data and the respective ensembles differ in the warming intensities in the 19th and 20th century for Europe and also in the last 100 yr for the Northern Hemisphere (Fig. 1a, b). For Europe, especially the strong solar forcing simulations differ in recent temperature evolutions. An over-estimation of variability is expected relative to the hemispheric mean reconstruction (Fig. 1e, see note in Sect. 2.3) but it also occurs with respect to an inter-annually representative South American temperature reconstruction (not shown). Nominally inter-annual standard deviations can be of comparable size in the reconstruction sub-ensemble and the target simulation ensemble mean (Fig. 1f). One reconstruction generally varies about twice as much as the simulated truth, while the true variability exceeds the variability of the reconstructions in periods of large volcanic eruptions (compare Fig. 1e, f, e.g. 13th, 15th and early 19th centuries, compare also Mann et al., 2012, and Briffa et al., 1998).

3.2 Spatial fields

3.2.1 Ensemble consistency of field estimates

In the following, the analyses of consistency are extended to the decadally smoothed global temperature field reconstruction by Mann et al. (2009). We note again that deviations from uniformity of the histograms may be due to deviations in one particular period, while other periods may display consistency between reconstructions and simulations. These discrepancies can easily be identified in the analysis of time series data. For the assessment of the spatial field data we consider the question of consistency at the grid-point level and do so for different time periods to highlight the possible deviations.

The reconstructed climatology for one part of the Little Ice Age period (1390s to 1690s) is displayed in Fig. 4a, and Fig. 4c shows the rank of the reconstruction data in the COSMOS-Mill ensemble of surface temperature data for this climatology. From Fig. 4b it can be seen that the simulations frequently vary more than the field reconstruction at individual grid points. Figures 5 to 7 display a selection of results for the evaluation of consistency. Although no uncertainty estimate is given for the global field data, we inflate the ensemble by a random error drawn from a distribution with a standard deviation equaling the largest standard error of the unscreened Northern Hemisphere mean temperature series provided by Mann et al. (2009). Without error inflation, expected effective rank frequencies can be very small considering the temporal autocorrelations in the data. The number of independent samples is always largest over the Tropical Pacific (not shown) probably due to the too strong and too regular ENSO in MPI-ESM (Jungclaus et al., 2006).

As for the time series data, the most common deviation is a too wide simulation ensemble for rank counts (Fig. 5 for a random selection of grid points) and residual quantiles (Fig. 6 for a random selection of grid points). However, the ensemble may arise as too narrow at individual grid points over the full period due to opposite probabilistic biases. Objectively flat rank counts are found as well for sub-periods and the

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

full period, although again opposite biases may lead to this result. The notable shifts in probabilistic consistency are highlighted by considering different periods of 250 records in the range from 805 to 1845 CE (Fig. 5). Outstanding changes occur between opposing biases, as the ensemble is found to be moderately (or even extremely) biased in at least one sub-period.

The prominent lack of consistency between simulations and the field reconstruction becomes even more obvious in the climatological residuals (Fig. 6). Among the individual ensemble members, the climatological behavior is mostly comparable relative to the reconstruction. The prominently sloped residual quantiles highlight the stronger variability in the ensemble even for decadal moving averages. However, at certain grid points under-dispersive or consistent climatologies can be seen. Changes in the r-q-q plots are diverse between periods but can be rather small between the first and the last 250 records (compare Fig. 6). Some improvement is seen towards more limited deviations or nearly vanishing residuals in the late period. At other grid points, biases increase, change sign or deviating spread characteristics become more pronounced. In compliance with the shifts in the probabilistic deviations, there are grid points where either the reconstructed quantile distributions or the anomaly quantile deviations or both are completely different between early and late records for the decadally smoothed global temperature data. Thus, results for sub-periods are often not comparable with each other in neither the probabilistic nor the climatogic evaluation. Occurring shifts emphasize the general lack of a common signal.

Decadal smoothing reduces the width of the climatological quantile distributions, and a number of grid points display only very small quantiles as a sign of very weak inter-decadal variability (not shown). At certain grid points, the extremely narrow reconstructed quantile distributions result in particularly strong climatological overdispersion. Quantile distributions are in parts broader in higher Northern Hemisphere latitudes for reconstructions and simulations.

The probabilistic consistency at each grid point of the global data is best visualized by displaying the results from the goodness-of-fit tests for the rank histograms. In Fig. 7

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

grid cells are colored with respect to the p-values of the goodness-of-fit test. Rejections of the uniform null hypothesis are displayed in red and p-values smaller than 0.1 in blue. The left column gives results for the general χ^2 test, the right displays the maximum of the p-values for single deviation tests for bias and spread. If no errors in the truth are considered (not shown), the full test generally does not reject uniformity for the full period. However, the single deviations are frequently significant especially over the oceans for the early and late periods of the data. Thus, while centering the data over the full period leads to consistent estimates from the late 11th to the early 16th century the long-term trends are notably different at the beginning and at the end.

If a moderate random error inflation is used, spatially extended consistency is mainly restricted, according to the full test, to Central Eurasia and the Tropical Pacific for the full period (Fig. 7a). For four sub-periods of 250 records diverging results become visible. For example, the pair of reconstruction and ensemble simulations is consistent in the North Atlantic sub-polar gyre region for the early period (Fig. 7b), but uniformity is rejected for the following 250 records (Fig. 7c). Overall, prominently opposite results arise in the full test for these early two periods, with wide regions of Eurasia and North America consistent in the latter but not in the early one. During the period from about 1300 to 1550 (the early Little Ice Age, Fig. 7d), the ensemble appears to be consistent in Northern North America, the Tropical Pacific and South of Greenland. In the last period (Fig. 7e, about 1595 to 1845), Eurasia and the North Atlantic again arise as the most consistent regions according to the full test including the uncertainty of the truth. On the other hand, single deviations are nearly always and everywhere significant (Fig. 7f–j). Deviations are least prominent close to the regions where the original proxy density was largest in the analysis of Mann et al. (2009).

If probabilistic and climatological consistency are assessed for all data points in space and time together, over-dispersion is again pronounced with respect to both aspects (not shown). The cumulative spatial assessment suggests strongly differing relations between reconstructed and simulated decadal temperatures on global scales (not shown).

In summary, as for the time series, the utilized simulation ensemble displays a lack of consistency with the global reconstruction. However, uniformity cannot be rejected for some regions and certain periods based on the full test, which may be to some extent due to a very small number of independent samples. The most prominent lack of consistency is seen over the southern oceans. Tests for the single deviations of bias and spread are nearly everywhere significant after inclusion of an error estimate. Thus, general consistency between simulations and reconstructions remains very weak. Note, (lack of) consistency is not homogeneous in time, but may differ between selected periods. The simple assumption of increasing consistency with decreasing temporal distance to the present is not necessarily valid.

3.2.2 Comparison of patterns and grid point variability of the spatial field reconstruction

Simulated mean anomalies seldom agree with reconstructed patterns for specific periods as can be inferred from the mapped ranks in Fig. 4c which refer to a sub-period of the Little Ice Age (1390s to 1690s). The reconstructed climatology map for this period is shown in Fig. 4a. While the amplitudes of mean anomalies are comparable between reconstructions and strong solar full-forcing simulations except in the Tropical Pacific, the weak solar full-forcing simulations display less cooling in the selected period (not shown, compare example map in Fig. 4a and rank map in Fig. 4c). Variability is as often comparable as not (Fig. 4b). The simulations especially vary more than the reconstruction over oceanic regions (middle blue in Fig. 4b). This relation is reverted over the Southern Hemisphere ocean, particularly the South Atlantic and in the Southern Indian ocean as seen in the relative standard deviations for the full period in Fig. 4b.

The ranks in Fig. 4c indicate a particularly strong and spatially extended mismatch between simulations and reconstructions in the tropical Pacific during the Little Ice Age. This strong signal is less due to the strong ENSO variability in MPI-ESM (compare Jungclaus et al., 2006), but more due to the contrast between the reconstructed mean warm anomaly and the diverse but generally much weaker simulated mean anomalies. The strong solar single and full-forcing simulations even display notable negative anomalies (not shown). We note that this La Niña-like response not only contrasts the results by Mann et al. (2009) but that such a La Niña signature during periods of solar forcing minima is further in contrast to the findings of Meehl et al. (2009) and Emile-Geay et al. (2007) studying, respectively, the effect of peak solar activity in the observed 11 yr cycle on the climate in the Pacific sector and the role of ENSO in the climate impact of changes in the solar forcing; see also the discussions by Misios and Schmidt (2012) on the relationship between solar insolation maxima and Tropical Pacific sea surface temperatures.

Generally, the spatially-resolved temperature reconstruction represents the largest absolute mean anomalies in the selected periods as seen in the mapped ranks in Fig. 4c. This holds also for other field reconstructions (not shown). It is most pronounced over the oceans for the decadally smoothed global data (Fig. 4c). Thus, either (i) the considered ensemble of simulations generally underestimates the size of the mean anomalies over the periods of interest with reconstructed warm anomalies being warmest and cold anomalies coldest, or (ii) the simulations vary notably more in the averaging periods, or (iii) the comparison between anomaly patterns are of reduced value due to a general dissimilarity between reconstructions and simulations. In the first two cases, the impression of over-dispersion results from a general misrepresentation of the mean climate.

In summing up, the simple comparison indicates limitations in the correspondence between simulated and reconstructed climate states, limitations that also encompass their variability. The assessment of the consistency on the other hand objectifies the comparison between simulations and reconstructions, and the goodness-of-fit test allows to summarise, in one Figure, the (dis-)agreement in terms of ensemble consistency.

4 Discussions of the results

Jungclaus et al. (2010) show good agreement between the full-forcing simulations in the COSMOS-Mill ensemble and the HadCRUT3v Northern Hemisphere temperature data for the 20th century, but they also highlight periods in which the simulations are rather warm compared to temperature reconstructions when anomalies are considered with respect to the period 1961–1990 (e.g. in the 12th and 13th centuries). Thus, the optimal case of comparable non-linear long-term trends is not given for the simulation ensemble and common reconstructions, and we have to account for differences in mean states by centering both estimates to a common period for the test of consistency (similar to traditional simulation-reconstruction comparisons, e.g. Jansen et al., 2007; Brázdil et al., 2010; Luterbacher et al., 2010; Jungclaus et al., 2010; Zorita et al., 2010; Zanchettin et al., 2012).

Further data sets: strong probabilistic and climatological deviations arise between the data presented above for the utilized uncertainty estimates, the reference periods and the non-smoothed hemispheric data. Results for the seasonal European temperature reconstructions by Luterbacher et al. (2002, 2004) and Xoplaki et al. (2005) and the South American austral summer temperature reconstructions by Neukom et al. (2011) confirm this generally over-dispersive character of the ensemble (not shown). We can generally reject uniformity at the grid point level and for area average series. Only the annual Central European temperature time series data arises as possibly fully consistent.

Consistency relative to individual Northern Hemisphere reconstructions: Sect. 3 only considers the ensemble mean of the Northern Hemisphere reconstruction ensemble (Frank et al., 2010), but even consistency of the single reconstructions with one another may be questioned. The reconstruction sub-ensemble recalibrated to 1920–1960 is consistent with respect to the recalibrated Moberg et al. (2005), Mann et al. (2008) and Juckes et al. (2007) reconstructions (not shown, no uncertainty inflation), but otherwise various deviations occur (not shown).

Consistency of simulation ensembles and individual Northern Hemisphere reconstructions: assessing pairs of simulation ensembles (all, weak, strong solar full-forcing) and single reconstructions (Frank et al., 2010, recalibrated to the 1920–1960 period, no uncertainty inflation), the simulation ensembles display least deviations relative to the data by Frank et al. (2007, for the full and the weak solar full-forcing ensembles) and Juckes et al. (2007, weak and strong solar full-forcing). The three-member strong solar full-forcing ensemble appears also to be consistent with the D'Arrigo et al. (2006), Briffa (2000), Hegerl et al. (2007) and Moberg et al. (2005) reconstructions.

Test of consistency for surrogate ensembles: surrogate simulation ensembles constructed from a long control-run are found to be consistent with an equivalent surrogate truth, one of the weak solar full-forcing simulations and the weak solar-only forcing simulation. The full test rejects uniformity in less than one percent of the 2201 surrogate ensembles. Spread and bias tests are significant for less than 50 tests. Thus, pairs of ensemble and truth appear to be generally consistent, if variability is restricted to the internal variability of the simulated system or variability that is only marginally different from the internal variability (compare Zanchettin et al., 2010). In line with similar considerations in seasonal and medium-range weather forecasting (Johnson and Bowler, 2009), ensembles are consistent as long as the true variability and the simulated variability are similar.

If the surrogates are assessed against the 521 members of the Frank et al. (2010) recalibration ensemble, about 20 % of the pairs arise as consistent with respect to the full test although they are objectively unrelated. Single spread test statistics are not significant in about 50 cases. Climatologically, the surrogate ensemble agrees better than the real ensemble with some members of the reconstruction sub-ensemble calibrated to 1920–1960, indicating strong deviations between forced reconstructed and simulated climate evolutions.

Further discussions: the only data that yields reasonable consistency with the simulation ensemble (the Central European temperature reconstructions by Dobrovolný et al., 2010) is an estimate for the last 500 yr and, therefore, may benefit from a more stable number of reliable available proxy indicators than longer period reconstructions. The forcing data for this period can also be assumed to be less uncertain compared to the full millennium. We remark that part of the large simulated climate variability is possibly due to the well known too strong and too regular El Niño variability in the considered climate simulator (Jungclaus et al., 2006) and the related teleconnections.

As noted in Sect. 2.3, it is convenient, but not necessarily appropriate to employ the raw ensemble reconstructions (Frank et al., 2010) as annually resolved data. Similarly, it is arguable whether an ensemble mean represents unfiltered annually resolved data. A posteriori, our approach seems to be valid for the comparison of the specific simulation ensemble mean with this particular reconstruction ensemble, but the larger variability in the simulations compromises the inverse consideration. Interestingly, the moving standard deviations of the ensemble means (simulations and reconstructions) evolve similarly in the period 1400 to 1900. The 20th century disagreement is possibly due to the evolution of the simulations with strong solar forcing.

With a focus similar to the approach utilized here, Hind et al. (2012) provide a statistical framework for assessing climate simulations against paleoclimate proxy reconstructions allowing for an irregular spatio-temporal distribution of proxy series. Their framework concentrates on the similarity between simulated and reconstructed series by analysing two newly developed correlation-based and distance-based test statistics. Hind et al. apply their approach in a pseudo-proxy experiment within the virtual reality of the COSMOS-Mill sub-ensembles to test for the distinguishability of the two sub-ensembles. They conclude that prior to drawing resilient conclusions from our model simulations we need more proxy series with high signal-to-noise ratios.

Finally, with more and more simulations becoming available, the CMIP5/PMIP3 ensemble of past1000-simulations (Taylor et al., 2012; Braconnot et al., 2012) offers the opportunity to evaluate our simulated and reconstructed knowledge in a multi-model context. Similarly, the PAGES 2K Network (http://www.pages-igbp.org/) aims to provide new regional reconstructions for all continental areas and the global ocean allowing a detailed assessment of the consistency of our two tools. Preliminary analyses for the available CMIP5/PMIP3-past1000-simulations indicate that the multi-model-ensemble behaves similar to the COSMOS-Mill ensemble with respect to probabilistic and climatological consistency relative to the European and northern hemispheric temperature reconstructions considered in the present manuscript.

5 Concluding remarks

Rank histograms, χ^2 goodness-of-fit test decomposition and residual quantile-quantile plots help to assess the probabilistic and climatological consistency of ensemble projections against an observed truth (e.g. Annan and Hargreaves, 2010). If no state of truth can be identified, as is the case in periods and regions without instrumental observations, such statistical analyses add an objective component to the evaluation of simulation ensembles and statistical approximations from paleo-sensor data (Braconnot et al., 2012) under uncertainty and beyond “wiggle matching”. The approach permits a succinct visualization of the consistency between an ensemble of estimates and an uncertain verification truth. Ideally, it also reduces the dependence on the reference climatology which is present in many visual and mathematical methods that aim to qualify the correspondence between simulations and (approximated) observations.

Considering the COSMOS-Mill-ensemble and various reconstructions within the described approach, we find the simulation ensemble to be consistent, within sampling variability, with the Central European temperature reconstruction by Dobrovolný et al. (2010). However, the ensemble lacks consistency with respect to the mean of the ensemble of Northern Hemisphere mean temperature reconstructions by Frank et al. (2010) due to probabilistic and climatological over-dispersion, as the ensemble samples from a significantly wider distribution than the reconstruction ensemble mean. The distribution of the reconstruction ensemble in turn is too wide relative to the simulation ensemble mean.

Similarly, the simulation ensemble is found to be statistically distinguishable from the global field temperature reconstruction by Mann et al. (2009). Although probabilistic consistency is found for multi-centennial sub-periods and certain regions according to the applied full test, accounting for single probabilistic deviations and climatological differences emphasizes a general lack of consistency. The largest, but still limited consistency is seen over areas of Eurasia and North America for both full and sub-periods. For some periods, we also cannot reject consistency for most tropical and northern hemispheric ocean regions. The profound lack of climatological and probabilistic consistency between the simulation ensembles and reconstructions stresses the importance of improving our two tools to investigate past climates in order to achieve a more resilient estimate of the truth.

If our estimates are not consistent with each other for certain periods and areas, it is unclear how we should compare their accuracy. Thus, if these reconstructions and these simulation ensembles are employed in dynamical comparisons and in studies on climate processes, we have to account for the climatological and probabilistic discrepancies between both data sets, that have been described in the present work.

Acknowledgements. Comments by Frank Sienz and James Annan on earlier versions of the manuscript lead to notable improvements; Madeleine Pascolini-Campbell and Mark Carson provided copy-editing. We thank the Model & Data group at MPI-M for providing the model data. This work benefited from the efforts of the Paleoclimatology Program at NOAA and its archive of reconstructions of past climatic conditions and forcings.

O. B. acknowledges funding through the Cluster of Excellence “CliSAP”, University of Hamburg, funded through the German Science Foundation (DFG). D. Z. was supported through the ENIGMA project of the Max Planck Society and from the Federal Ministry for Education and Research in Germany (BMBF) through the research program “MiKlip” (FKZ:01LP1158A).

This work has been carried out as part of the MPI-M Integrated Project Millennium and contributes to the Cluster of Excellence “CliSAP” at Hamburg University.

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

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996. 2412, 2413, 2414

Annan, J. D. and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, doi:10.1029/2009GL041994, 2010. 2412, 2413, 2414, 2430

Annan, J. D., Hargreaves, J. C., and Tachiiri, K.: On the observational assessment of climate model performance, Geophys. Res. Lett., 38, L24702, doi:10.1029/2011GL049812, 2011. 2411

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, J.-Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, Th., Hewitt, C. D., Kageyama, M., Kitoh, A., Laîné, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S. L., Yu, Y., and Zhao, Y.: Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features, Clim. Past, 3, 261–277, doi:10.5194/cp-3-261-2007, 2007. 2412

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. 2411, 2420, 2429, 2430

Bradley, D. R., Bradley, T. D., McGrath, S. G., and Cutcomb, S. D.: Type I error rate of the Chi-square test in independence in $R \times C$ tables that have small expected frequencies, Psychol. Bull., 86, 1290–1297, doi:10.1037/0033-2909.86.6.1290, 1979. 2414

Bradley, R. S.: High-resolution paleoclimatology, in: Dendroclimatology, edited by: Hughes, M. K., Swetnam, T. W., and Diaz, H. F., Developments in Paleoenvironmental Research, volume 11, chapter 1, Springer, Dordrecht, 3–15, doi:10.1007/978-1-4020-5725-0_1, 2011. 2411

Brázdil, R., Dobrovolný, P., Luterbacher, J., Moberg, A., Pfister, C., Wheeler, D., and Zorita, E.: European climate of the past 500 years: new challenges for historical climatology, Climatic Change, 101, 7–40, doi:10.1007/s10584-009-9783-z, 2010. 2427

Briffa, K. R.: Annual climate variability in the holocene: interpreting the message of ancient trees, Quat. Sci. Rev., 19, 87–105, doi:10.1016/S0277-3791(99)00056-6, 2000. 2428

Briffa, K. R., Jones, P. D., Schweingruber, F. H., and Osborn, T. J.: Influence of volcanic eruptions on Northern Hemisphere summer temperature over the past 600 years, Nature, 393, 450–455, doi:10.1038/30943, 1998. 2421

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. 2411

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. 2428

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. 2412, 2415, 2418, 2428, 2430

Efron, B. and Tibshirani, R. J.: An Introduction to the Bootstrap, Monographs on Statistics & Applied Probability, Chapman and Hall/CRC, New York, 1st Edn., 1994. 2414

Emile-Geay, J., Cane, M., Seager, R., Kaplan, A., and Almasi, P.: El Niño as a mediator of the solar influence on climate, Paleoceanography, 22, PA3210, doi:10.1029/2006PA001304, 2007. 2426

Frank, D., Esper, J., and Cook, E. R.: Adjustment for proxy number and coherence in a large-scale temperature reconstruction, Geophys. Res. Lett., 34, L16709, doi:10.1029/2007GL030571, 2007. 2428

Frank, D. C., Esper, J., Raible, C. C., Büntgen, U., Trouet, V., Stocker, B., and Joos, F.: Ensemble reconstruction constraints on the global carbon cycle sensitivity to climate, Nature, 463, 527–530, doi:10.1038/nature08769, 2010. 2411, 2415, 2427, 2428, 2429, 2430

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., 113, D23111, doi:10.1029/2008JD010239, 2008. 2411

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001. 2413, 2416

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. 2412, 2413

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. 2428

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, Clim. Past Discuss., 8, 263–320, doi:10.5194/cpd-8-263-2012, 2012. 2429

Jansen, E., Overpeck, J., Briffa, K. R., Duplessy, J. C., Joos, F., Masson-Delmotte, V., Olago, D., Otto-Bliesner, B., Peltier, W. R., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R., and Zhang, D.: Palaeoclimate, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 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 and New York, NY, USA, 2007. 2427

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. 2428

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. 2412, 2413, 2414

Joussaume, S. and Taylor, K. E.: The paleoclimate modelling intercomparison project, in: Paleoclimate Modelling Intercomparison Project (PMIP): Proceedings of the Third PMIP Workshop, edited by: Braconnot, P., Canada, 43–50, 2000. 2412

Juckes, M. N., Allen, M. R., Briffa, K. R., Esper, J., Hegerl, G. C., Moberg, A., Osborn, T. J., and Weber, S. L.: Millennial temperature reconstruction intercomparison and evaluation, Clim. Past, 3, 591–609, doi:10.5194/cp-3-591-2007, 2007. 2427, 2428

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, doi:10.1175/JCLI3827.1, 2006. 2422, 2425, 2429

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. 2411, 2415, 2427

Luterbacher, J., Xoplaki, E., Dietrich, D., Rickli, R., Jacobeit, J., Beck, C., Gyalistras, D., Schmutz, C., and Wanner, H.: Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500, Clim. Dynam., 18, 545–561, doi:10.1007/s00382-001-0196-6, 2002. 2427

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European seasonal and annual temperature variability, trends, and extremes since 1500, Science, 303, 1499–1503, doi:10.1126/science.1093877, 2004. 2427

Luterbacher, J., Koenig, S. J., Franke, J., Schrier, G., Zorita, E., Moberg, A., Jacobeit, J., Della-Marta, P. M., Küttel, M., Xoplaki, E., Wheeler, D., Rutishauser, T., Stössel, M., Wanner, H., Brázdil, R., Dobrovolný, P., Camuffo, D., Bertolin, C., Engelen, A., Gonzalez-Rouco, F. J., Wilson, R., Pfister, C., Limanówka, D., Nordli, Leijonhufvud, L., Söderberg, J., Allan, R., Barriendos, M., Glaser, R., Riemann, D., Hao, Z., and Zerefos, C. S.: Circulation dynamics and its influence on European and Mediterranean January–April climate over the past half millennium: results and insights from instrumental data, documentary evidence and coupled climate models, Climatic Change, 101, 201–234, doi:10.1007/s10584-009-9782-0, 2010. 2427

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, Proc. Natl. Acad. Sci., 105, 13252–13257, doi:10.1073/pnas.0805721105, 2008. 2427

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. 2412, 2414, 2415, 2417, 2422, 2424, 2426, 2430

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. 2421

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, 2010. 2412, 2413, 2414, 2415, 2416

Meehl, G. A., Arblaster, J. M., Matthes, K., Sassi, F., and van Loon, H.: Amplifying the Pacific climate system response to a small 11-year solar cycle forcing, Science, 325, 1114–1118, doi:10.1126/science.1172872, 2009. 2426

Misios, S. and Schmidt, H.: Mechanisms involved in the amplification of the 11-yr solar cycle signal in the Tropical Pacific Ocean, J. Climate, in press, doi:10.1175/JCLI-D-11-00261.1, 2012. 2426

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlen, W.: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433, 613–617, doi:10.1038/nature03265, 2005. 2427, 2428

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteorol., 12, 595–600, 1973. 2413

Neukom, R., Luterbacher, J., Villalba, R., Küttel, M., Frank, D., Jones, P. D., Grosjean, M., Wanner, H., Aravena, J. C., Black, D. E., Christie, D. A., D'Arrigo, R., Lara, A., Morales, M., Soliz-Gamboa, C., Srur, A., Urrutia, R., and Gunten, L.: Multiproxy summer and winter surface air temperature field reconstructions for Southern South America covering the past centuries, Clim. Dynam., 37, 35–51, doi:10.1007/s00382-010-0793-3, 2011. 2427

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., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007. 2411

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. 2411

Schrijver, C. J., Livingston, W. C., Woods, T. N., and Mewaldt, R. A.: The minimal solar activity in 2008–2009 and its implications for long-term climate modeling, Geophys. Res. Lett., 38, L06701, doi:10.1029/2011GL046658, 2011. 2411

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. 2411

Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, Geophys. Res. Lett., 36, L19704, doi:10.1029/2009GL040142, 2009. 2411

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bull. Am. Meteorol. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012. 2429

Toth, Z., Talagrand, O., Candille, G., and Zhu, Y.: Probability and ensemble forecasts, in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by: Jolliffe, I. T. and Stephenson, D. B., John Wiley, Chichester, UK, 137–163, 2003. 2412

Wilks, D. S.: On the reliability of the rank histogram, Mon. Weather Rev., 139, 311–316, doi:10.1175/2010MWR3446.1, 2010. 2412, 2414

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., 112, D17103, doi:10.1029/2006JD008318, 2007. 2411

Xoplaki, E., Luterbacher, J., Paeth, H., Dietrich, D., Steiner, N., Grosjean, M., and Wanner, H.: European spring and autumn temperature variability and change of extremes over the last half millennium, Geophys. Res. Lett., 32, L15713, doi:10.1029/2005GL023424, 2005. 2427

Zanchettin, D., Rubino, A., and Jungclaus, J. H.: Intermittent multidecadal-to-centennial fluctuations dominate global temperature evolution over the last millennium, Geophys. Res. Lett., 37, L14702, doi:10.1029/2010GL043717, 2010. 2428

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, Clim. Dynam., in press, 1–18, doi:10.1007/s00382-012-1361-9, 2012. 2427

Zorita, E., Moberg, A., Leijonhufvud, L., Wilson, R., Brázdil, R., Dobrovolný, P., Luterbacher, J., Böhm, R., Pfister, C., Riemann, D., Glaser, R., Söderberg, J., and González-Rouco, F.: European temperature records of the past five centuries based on documentary/instrumental information compared to climate simulations, Climatic Change, 101, 143–168, doi:10.1007/s10584-010-9824-7, 2010. 2427

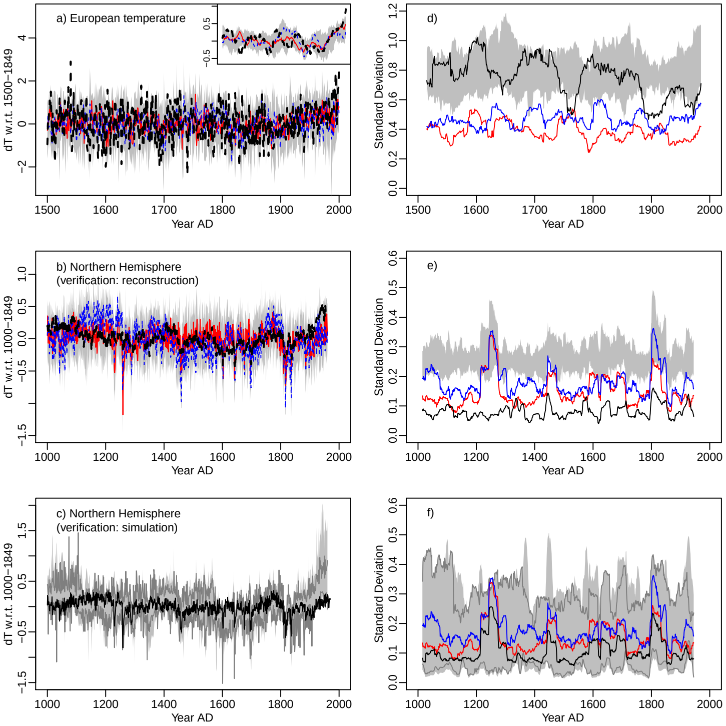

Fig. 1. (a–c) Time series. **(d–f)** Moving 31-yr standard deviations. **(a, d)** European annual temperature, **(b, e)** Northern Hemisphere simulation ensemble against reconstructed truth, **(c, f)** Northern Hemisphere reconstruction ensemble versus simulated truth. In all panels, black is the respective verification data and grey shading is the range of the ensembles. In **(a, b, d, e, f)** red (blue) lines are the weak (strong) solar full-forcing simulation ensemble means. In **(c, f)** the range of the reconstruction sub-ensemble recalibrated to the period 1920 to 1960 is displayed in grey lines. Inset in **(a)** presents the 31-yr moving averages of the European estimates, and we choose to present the truth and the strong solar full-forcing simulation ensemble means in **(a, b)** by dashed lines to increase the visibility of all time series.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀I ▶I
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

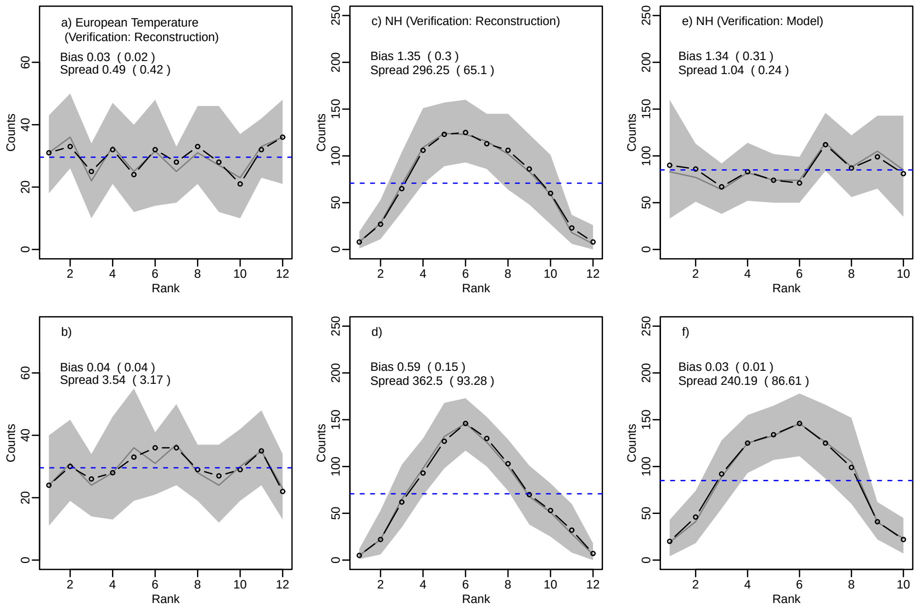

Fig. 2. Rank histogram counts for temperature data including **(a, b)** Central European annual temperatures, **(c, d)** Northern Hemisphere simulation ensemble temperature, **(e, f)** Northern Hemisphere reconstruction sub-ensemble calibrated to the period 1920 to 1960. Top (bottom) row does (does not) neglect the errors in the truth. Numbers are χ^2 statistics. χ^2 statistics in brackets account for auto-correlation in the data. Grey shading (line) are 0.5 % and 99.5 % (50 %) quantiles for block-bootstrapped rank histograms (2000 replicates, block length of 50 yr). Blue horizontal lines give the expected average count for a perfectly uniform histogram.

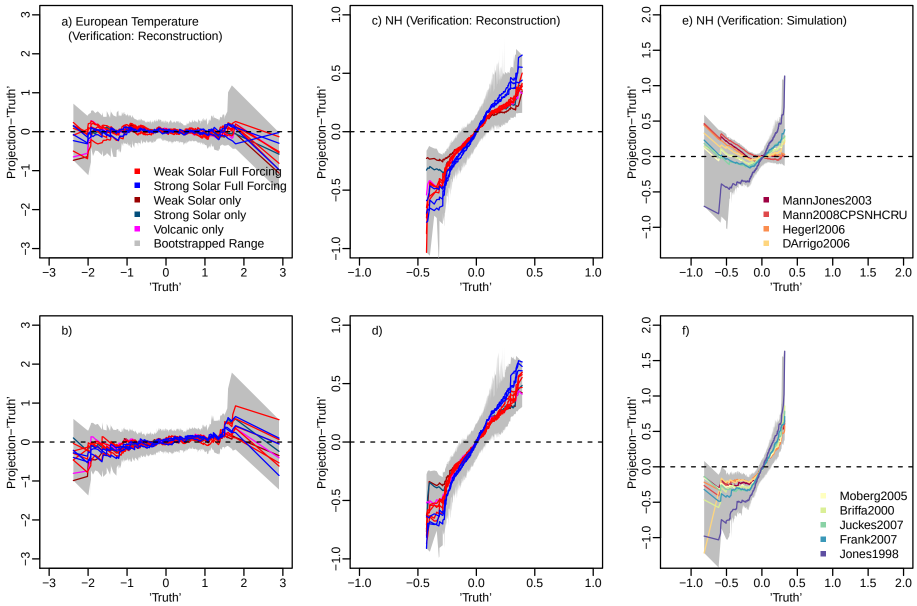

Fig. 3. Residual quantile-quantile plots for temperature data including **(a, b)** Central European annual temperatures, **(c, d)** Northern Hemisphere simulation ensemble temperature, **(e, f)** Northern Hemisphere reconstruction sub-ensemble calibrated to the period 1920 to 1960. Top (bottom) row does (does not) neglect the errors in the truth. See legend for individual ensemble members. Grey shading are 0.5 % and 99.5 % quantiles for block-bootstrapped residual quantiles (2000 replicates, block length of 50 yr).

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

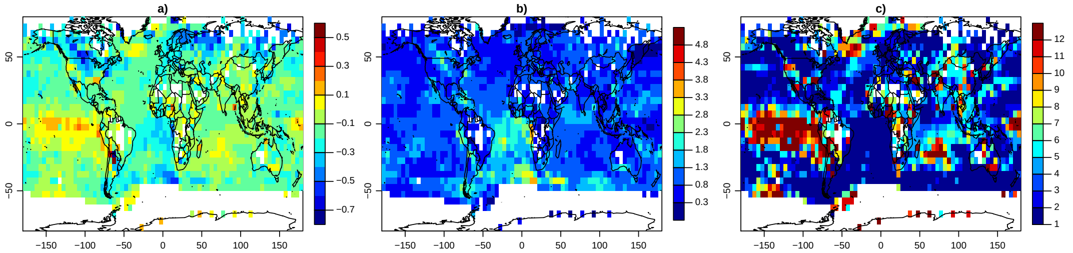

Fig. 4. Global decadal smooth temperature: **(a)** reconstructed mean anomaly map for a cold period (for the 1390s to 1690s), **(b)** ensemble mean of relative standard deviations (reconstruction standard deviation divided by simulation standard deviation at each grid point for the full period), **(c)** mapped ranks for the cold period (1390s to 1690s).

CPD

8, 2409–2444, 2012

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

|◀ | ▶|

◀ | ▶

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

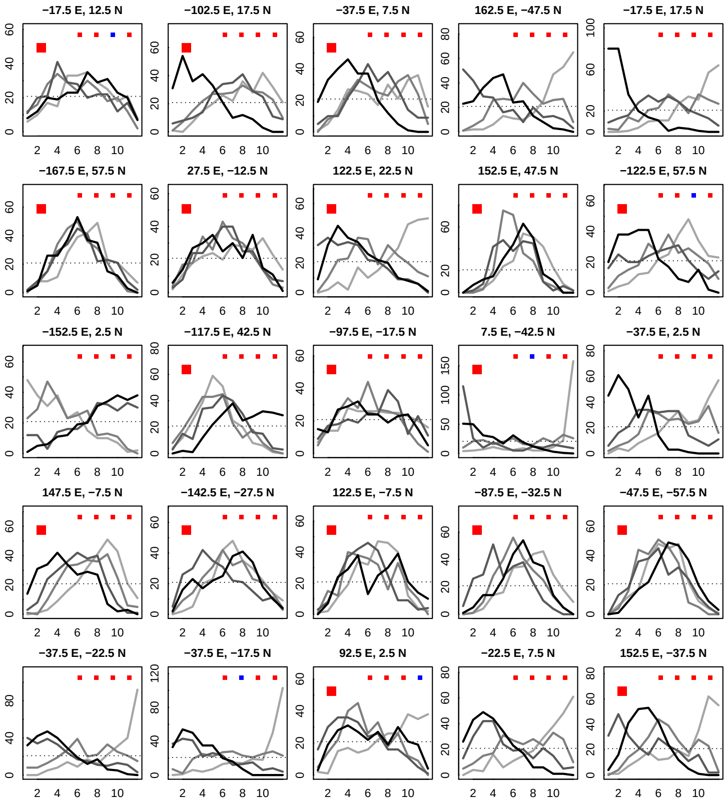

Fig. 5. Rank histogram counts for a random selection of 25 grid points from the decadal smooth global temperature data and the first, second, third and last 250 records of the decadally smoothed annual data (grey to black lines). Large (small) red squares mark grid points where spread or bias deviations are significant over the full (the individual sub-)period. Blue squares are not significant. Squares from left to right for the first, second, third and last sub-period. Locations given in titles of individual panels.

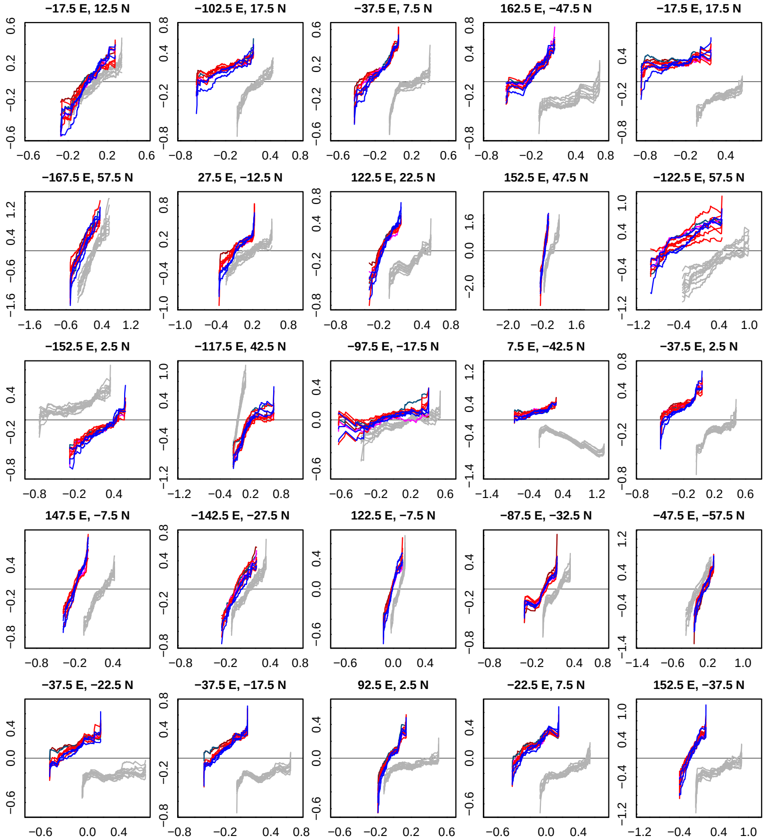

Fig. 6. Residual quantile-quantile plots for a random selection of 25 grid points from the decadal smooth global temperature data and the first (grey) and the last (colors) 250 records. Locations given in titles of individual panels. Representation as in Fig. 3a.

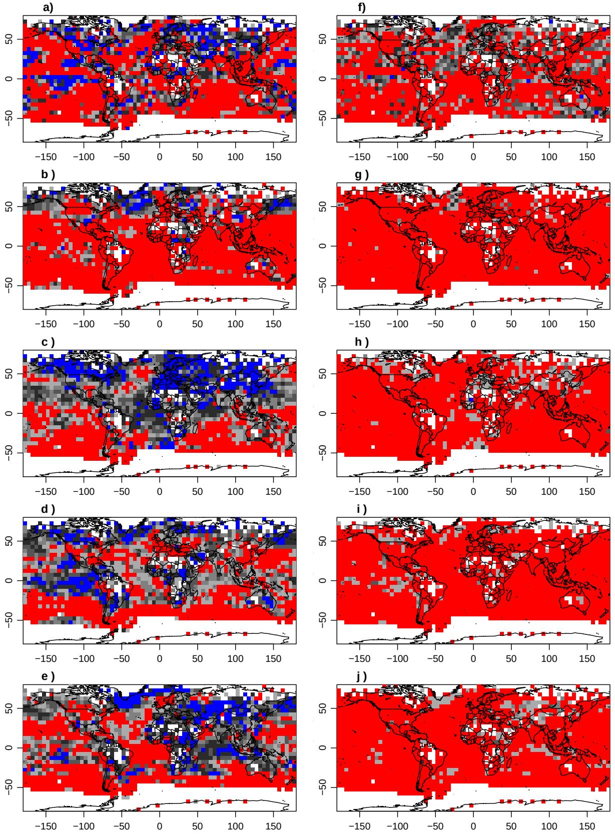

Fig. 7. Global assessment of the goodness-of-fit test for the decadal smooth data considering errors in the truth. Plotted are lower p-values. In the left column: full χ^2 test, in the right column: maximum of p-values for single deviation tests for bias and spread. Blue smaller 0.1, dark to light grey in steps of 0.2 the range between 0.1 and 0.9, red larger than 0.9. **(a, f)** full period and **(b–e)** and **(g–j)** for the first, second, third and last period of 250 records.