Manuscript prepared for J. Name
with version 4.2 of the LaTeX class copernicus.cls.
Date: 16 November 2012

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

Oliver Bothe[1,2,5], Johann H. Jungclaus[1], Davide Zanchettin[1], and Eduardo Zorita[3,4]

[1]Max-Planck-Institute for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany
[2]University of Hamburg, KlimaCampus Hamburg
[3]Institute for Coastal Research, Helmholtz Centre Geesthacht
[4]Bert Bolin Centre for Climate Research, University of Stockholm
[5]now at Leibniz Institute for Atmospheric Physics e.V. at the University of Rostock

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

Abstract. Are simulations and reconstructions of past climate and its variability comparable with each other? We assess ~~if~~ whether simulations and reconstructions for the climate of the last millennium 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~~ serve as verification data for a set of 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~~ often too wide relative to the employed reconstructions. ~~Similarly, an ensemble of reconstructions is too wide when evaluated against~~ An ensemble of Northern Hemisphere reconstructions is possibly consistent with the simulation ensemble mean ~~.~~

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

~~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~~ If we treat simulations and reconstructions as equitable hypotheses about past climate variability, a lack of their consistency weakens our confidence in inferences about past 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. Therefore a univocal estimation of accuracy is not possible if we acknowledge that our available estimates of past climate evolutions are on an equal footing but, as shown here, inconsistent with each other.

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 on a past climate state requires reconciling both estimatesin terms of accuracy and reliability. In case of an ensemble of estimates, we have . As a first step we may apply methods from numerical weather forecast verification (see, e.g., Toth et al., 2003; Marzban et al., 2010; Persson, 2011) to evaluate the consistency of the ensemble with relevant validation an ensemble of estimates with relevant verification data.

Similar to measurements by instrumental sensors, our Our pseudo-observations by proxies or paleo-sensors (as coined by Braconnot et al., 2012) are subject to "measurement" uncertainty similar to measurements by instrumental sensors. Uncertainties enter our reconstructions, among other ways, through the dating of the non-climate observationpseudo-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, e.g., 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 exampleThus, we have no independent direct and reliable observational knowledge on the climate 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" . We can increase our confidence in the two tools by applying (weather-)forecast verification methods which are less subjective than by-eye evaluations. Practically we select a verification data target from the available reconstructions (simulations)to verify an available ensemble of simulations, and vice versa. For a –specific task at hand, the analysis of consistency identifies whether the simulated and reconstructed climate estimates ensemble and the verification target can be consid-

ered to be compatible ~~realizations of an unknown "true" distribution, though not necessarily identical with it~~ realisations of the unknown past climatology or the unknown past climate evolution. Such compatibility has to consider probabilistic and climatological properties. Note, to be compatible does not imply to be identical (see for example Annan et al., 2011). Here we consider compatibility in terms of the concept of ensemble consistency as used in the field of weather-forecast verification (e.g. Marzban et al., 2010). Reconstructions and simulations are therefore treated as different but equitable hypotheses~~. Ensembles of hypotheses are available for northern hemispheric mean temperature reconstructions and for the PMIP3-compliant Community Simulations of the last millennium allowing the assessment of the consistency of reconstructions and simulations~~ , and there consistency is assessed 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~~ consistency 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~~ Northern Hemisphere mean temperature for the last millennium. Relevant ensembles are available for reconstructions (Frank et al., 2010) and for the PMIP3-compliant Community Simulations of the last millennium (COSMOS-Mill, Jungclaus et al., 2010). 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~~ the assessment of mean-state quantities (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~~ goodness-of-fit test-statistic (Jolliffe and Primo, 2008). The ~~restrictions of the approach are considered~~ climatological component of ensemble consistency is evaluated by presenting residual quantile-quantile plots (Marzban et al., 2010; Wilks, 2010)~~to evaluate the climatological consistency~~. The methods are discussed in Sect. 2 which further introduces the necessary terminology. Section 3 first discusses the robustness of the approach and then presents results concerning the consistency of reconstructions and simulations~~, and the sensitivity of the chosen approach is discussed in Sect. 4.~~ .

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~~ a suitable verification data target or by evaluating the ~~reliability~~ consistency of the full ensemble ~~, that is the compliance between "true" and ensemble estimated probability distributions . Considering the multiple sources of uncertainties in paleo-climate reconstructions and simulations, assessing ensemble consistency objectifies~~ (e.g. Marzban et al., 2010). We use the term "consistency" in the sense as used by, e.g. Annan and Hargreaves (2010) and Marzban et al. (2010). This usage is similar to verifying the reliability of numerical weather forecast ensembles. Thus, consistency refers to the agreement between the frequencies of the ensemble estimated data and the verification target. We assume an ensemble to be consistent if we cannot reject the hypothesis that the frequencies of occurrence of events in the ensemble are equal to the frequencies in the verification data. Since large uncertainties are associated with our data and we lack an observed verification target, the assessment of ensemble consistency provides a necessary condition for our evaluation of ensemble accuracy ~~. In the following, if we mention a "truth" or a "true" dataset, this can only represent an uncertain approximation of the observable truth~~in paleoclimate-studies (following Annan and Hargreaves, 2010).

There are two components of consistency to be considered, probabilistic and climatological consistency (Johnson and Bowler, 2009; Marzban et al., 2010). That is, we have not only to evaluate whether within-ensemble frequencies are consistent with those of the verification data, but also whether the variance of the ensemble member climatologies agree with the verification climatology.

The ~~reliability of a probabilistic ensemble is commonly validated~~ probabilistic component is commonly evaluated under the paradigm of statistical indistinguishability by ranking ~~true observational~~ the verification target data against the ensemble data (Anderson, 1996; Jolliffe and Primo, 2008; Annan and Hargreaves, 2010; Marzban et al., 2010; Hargreaves et al., 2011). ~~True~~ Target data 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~~ Indistinguishability refers to the assumption that the verification data may be exchanged for any member of the ensemble without changing the characteristics of the ensemble. Verification target and ensemble estimated (e.g. forecasted) frequencies agree for a consistent (or reliable) ensemble (Murphy, 1973). That is, we expect equiprobable outcomes for an ideal ensemble, and the ranking should result in a uniform, flat histogram. ~~For a "reliable" ensemble , observed and ensemble estimated (e.g. forecasted) frequencies agree . Note , however,~~ The underlying null hypothesis is of an overarching distribution for verification and ensemble data; for any data point, the ensemble and the verification data are assumed to originate from distributions which are similar enough to be

indistinguishable. Note that a flat histogram of ranks does not ~~necessarily imply reliability~~ imply consistency (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;~~ test is suitable to test for the goodness of fit of the rank histogram relative to the flat expectation. Furthermore, Jolliffe and Primo (2008) ~~present a comprehensive delineation. In mapping spatial fields of verification ranks for climatological periods of interest (Sects. 3.3.1 and 3.3.2), 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~~present a decomposition of the test statistic which enables tests for individual deviations from flatness resulting from biases or from different widths of the distributions; see Jolliffe and Primo (2008) for a comprehensive delineation. Thus, we are able to test the consistency of an ensemble by considering the goodness-of-fit χ^2 statistics and the respective p-values dependent on the degrees of freedom of the distribution (e.g. Jolliffe and Primo, 2008). Distributional degrees of freedom equal $n-1$ for the full test (with n being the number of classes in the rank histogram) and 1 for the single deviation test (Jolliffe and Primo, 2008; Annan and Hargreaves, 2010). We reject consistency for certain right p-values of the test. However, we also interpret the test statistics in terms of a reversed null hypothesis, where appropriate, to test the hypothesis that there is a deviation from uniformity. This refers to the general goodness-of-fit χ^2 statistic or to a specific deviation for the decomposed statistic. It is reasonable to consider significance at a conservative one-sided 90% level due to the large uncertainties associated with the data. Thus, critical χ^2 values become 2.706 for the single deviation test. For the full test for consistency, we subsequently are going to consider ensembles of eleven, nine, five and three members (see Section 2.2). Critical values are respectively 17.275, 14.684, 9.236 and 6.251.

~~Meaningful statistics require to account~~ Marzban et al. (2010, see also Wilks, 2010) recommend to evaluate the climatological component of consistency using residual quantile-quantile plots (r-q-q plots). These are similar to common quantile-quantile plots since they also evaluate estimates for the climatological quantiles for the ensemble members against the verification data quantiles. Here, however, the differences between the simulated distribution quantiles and the verification data quantiles are displayed to emphasise deviations in the climatological distributions. These deviations include, among others, differences in the tails, the skewness and the mean of the distributions.

The visualization of estimated quantiles against a theoretical quantile distribution allows to assess whether the generating process for the estimates is of a similar structure as the theoretically assumed

process. Thus, for the present consideration, quantiles allow to identify whether for each individual ensemble member the structure of the (empirical) cumulative probability distribution agrees with that of the verification data sample. Plotting the residuals of the estimated quantiles eases the interpretation since agreement of estimated and verification quantiles is signalled by vanishing residuals.

A number of further assumptions enter our analyses. First, meaningful statistics require accounting for dependencies in the data (Jolliffe and Primo, 2008; Annan and Hargreaves, 2010) by e.g. evaluating the effective ~~degrees of freedom~~ size of independent samples in the time series. ~~A higher number of degrees of freedom~~ All analyses account for effective sample-size (see discussions by and references of Bretherton et al., 1999). Nevertheless, the results are sensitive to the assumptions made especially with respect to the included uncertainty estimates (see below in this section). A larger effective sample size 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~~ Temporal correlations in the data may further affect the structure of the rank histograms in the assessment of the consistency of time series (Marzban et al., 2010; Wilks, 2010)~~. Accounting for the sampling variability reduces the risk of drawing~~. Sampling variability can result in 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~~. This also allows for a secondary test of uniformity. The results are sensitive to the chosen block length and 50 yr are possibly too short according to the auto-correlation functions for some reconstructions. However, 50 yr appear to be a reasonable compromise if we consider that the optimal length may also be shorter for some records.

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.

~~recommend to evaluate the climatological component of reliability using residual quantile-quantile plots (~~Both methods, rank histograms and 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 distributionquantiles and the true quantiles emphasizes deviationsin the climatological distributions. Biases result in a horizontal displacement from zero in the~~ plots visually assist in identifying discrepancies between the ensemble data and the verification data. For the probabilistic assessment, a rank histogram of apparent dome-shape (u-shape) indicates that the verification data is sampling from a distribution narrower (wider) than the ensemble, thus the spread of the ensemble is overly wide (too narrow). 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.

Consequently, a rank histogram with substantial clustering on one end of the counts suggests that the ensemble data is less likely to come from the high (negative bias) or small classes (positive bias) of the distribution. The dome- or u-shapes, on the other hand, signify that the ensemble data has, respectively, a larger or smaller variance compared to the verification data. As discussed by, e.g., Jolliffe and Primo (2008), there are other possible deviations, but we concentrate on these two.

In the climatological r-q-q ~~plots, and climatological~~ plots, differences in the mean of the distributions, that is biases, are seen as horizontal displacements from the expectation of vanishing quantiles. Climatological over- and under-dispersion (too wide or too narrow distributions) relate to positive or negative slopes (Marzban et al., 2010). That is, if an individual ensemble member features a larger climatological variance than the verification data, a positive slope in the residual quantile occurs since the ensemble data systematically overestimates the distance between the mean and the quantile locations. Smaller climatological variance results in a negative slope since the quantiles are closer to the mean. Marzban et al. (2010) give more details on the interpretation of the pattern of residual quantiles.

Thus, the rank histograms are a tool to disclose whether a probabilistically interpreted ensemble and its verification data represent different climates. Quantiles or residual quantiles complement the analysis to account for differences in the climatologies of the ensemble members. In climate studies, they are especially able to highlight differences in the resolved values close to the tails. The rank histograms provide a means for evaluating the consistency of the joint distribution for the ensemble and verification data (see Wilks, 2010), and the residual quantiles highlight deviations between climatologies for individual simulations.

In addition, the ranking of the verification against the ensemble assists in evaluating gridded spatial data. Therefore, the position of the verification data within the ensemble can be visualized in maps (Sects. 3.3.1 and 3.3.2). At each grid-point the rank of the verification data is plotted for individual time steps or climatological periods. Individual low ranks of the target hint to a localised overestimation of the climate parameter by the ensemble in such spatially mapped verification ranks, whereas high ranks imply a negative bias in the simulation ensemble.

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 the atmosphere model ECHAM5, the ocean model 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 so-

lar 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. Additionally, we take advantage of the 3100-year control run describing an unperturbed climate.

~~Considered reconstructions are~~ All reconstructions are for annual mean temperature. We use 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 data , we~~ Whereas all data have an annual resolution, some are temporally smoothed (e.g. Mann et al., 2009). We reverse the approach for the Frank et al. (2010) data to study additionally the consistency of a reconstruction ~~sub-ensemble~~ ensemble with respect to ~~the~~ a simulation ensemble mean~~; we~~. Note that such an ensemble of reconstructions is only available for the hemispheric mean data. Frank et al. (2010) recalibrate the reconstructions by Jones et al. (1998), Briffa (2000), Mann and Jones (2003), Moberg et al. (2005), D'Arrigo et al. (2006), Hegerl et al. (2007), Frank et al. (2007), Juckes et al. (2007) and Mann et al. (2008) to various periods of instrumental observations. The last available annual data differ among the reconstructions considered by Frank et al. (2010). We use the sub-ensemble ~~calibrated~~ re-calibrated to the period 1920 to ~~1960.~~ 1960 for our reverse-analysis (in the following just referred to as sub-ensemble; see Frank et al., 2010, for discussion on the ensemble construction). This period likely presents the most reliable observational data if we want to use all nine reconstructions. The choice of the calibration window strongly influences the variability of the reconstructions which is going to influence the assessment of consistency. Assumed uncertainties generally base on the full ensemble and therefore should account for this sensitivity.

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. Additionally, four sub-periods are considered for the global field data consisting of non-overlapping 250 records. The first three periods cover the first 750 records of the full data and the last period covers the last 250 records of the data sets. Thus there is a gap between the first three and the last sub-period.

For the Central European data, the uncertainty is sampled from a normal distribution with zero mean and standard deviation equal to the one standard error estimate given by Dobrovolný et al.

(2010). For an ensemble of data, an observational uncertainty can be randomly sampled from a distribution with zero mean and standard deviation equal to the ensemble standard deviation at each point. For the ensemble-mean approach, we additionally use additive internal variability estimates for the target data (see Sect. 2.3 for details). No uncertainty estimate is given for the global field data. We choose to inflate the ensemble by a random uncertainty estimate drawn from a Gaussian distribution with standard deviation equal to the largest standard error (i.e. $\sigma = 0.1729$) of the unscreened Northern Hemisphere mean temperature series provided by Mann et al. (2009).

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~~ covariability in the verification data.

Already the first applications of the rank histogram advised caution in the interpretation (e.g. Anderson, 1996) not least because of the uncertainties in the verification data. More recently Hamill (2001), Marzban et al. (2010) and others discussed the influence of, e.g., the underlying distributions or temporal correlations on the results; see also Wilks (2010) and the references in these publications. Marzban et al. (2010) further discuss the influence of intra-ensemble correlations and correlations between the ensemble and the verification on the rank histogram.

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~~ consistency (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~~ The problem of sampling variability may also affect the evaluation of the climatological anomaly distributions. The amount of correlations between verification and ensemble or within the ensemble and the differences between both can result in misleading rank histograms under idealised conditions (see Marzban et al., 2010, for details). We do not discuss this effect here. However, we note that the intra-ensemble-correlations do not allow a priori to exclude a uniform outcome, while the ensemble-verification-correlations suggest that we may expect a u-shaped rank-count for some cases. These expectations are made under idealised assumptions which do not necessarily hold for the considered ensembles. We do not perform sensitivity analyses on how the correlations may affect the results for the considered ensembles. We assume that these caveats increase the general uncertainty in the comparison between simulations and reconstructions

of past climate states and variability. At this point we stress that we see consistency as a necessary condition for the comparability between simulations and reconstructions. Caveats and inconsistencies have to be considered in subsequent analyses of the simulated and reconstructed data.

~~Although the~~ The data sets are a priori assumed to represent ~~annually resolved values, this~~ inter-annual variations of the data. This is not necessarily valid~~. If the target /truth~~ , if the target is an ensemble mean, ~~the target displays reduced inter-annual~~ since this is going to display reduced variability compared to the ensemble members ~~. This has to be taken into account in interpreting the results~~especially on the inter-annual time scale. It is therefore likely that using an ensemble mean as ~~truth~~ verification data will change the ensemble consistency. Considering an ~~error in the truth can compensate such problems~~uncertainty estimate in the target can compensate this. If reconstruction and simulation ensemble estimates are thought to include the same ~~externally forced~~ externally-forced variability, the ~~true~~ target ensemble mean should essentially recover the forced signal within the propagated uncertainties, and the probabilistic ensemble estimates (including the uncertainty of the ~~truth~~target) should reliably represent the ~~true distribution. Similarly, members~~ target distribution. However, an alternative approach to compensate for the reduced variability is to add an estimate of the internal variability to the ensemble mean estimate. In the following we pursue this approach. Thus, for the evaluation of the simulation ensemble, we fit autoregressive-moving-average models to the residual deviations of the full reconstruction ensemble from the ensemble mean. Thereby we obtain 521 possible fits. We produce for each fit 10 random representations of the process to add to the ensemble mean. For the reverse analyses, we add one section of the control run (Jungclaus et al., 2010) to the ensemble mean simulation. We regard using only one segment robust enough for evaluating the internal variability of the simulations since we further account for the sampling variability.

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~~ our assessment, we test for the consistency of the rank histograms for our ensembles with the hypothesis of a uniform outcome. We start by looking at the intra-ensemble consistency before assessing the area-averaged and field estimates.

3.1 Intra-ensemble consistency

Before we evaluate the consistency of the chosen simulation and reconstruction data sets, it is in place to describe the within-ensemble consistencies. We construct a surrogate simulation ensemble from the control-run. This ensemble is found to be probabilistically consistent with 2201 equivalent surrogate targets, three of the weak solar full-forcing simulations, the weak solar-forcing only simulation and the volcanic-forcing only simulation. The full test rejects uniformity in less than one percent of the tested surrogate targets (see Figure 1a). Spread and bias tests are significant for less than 50 tests (see Figure 1a). Here, we do not include uncertainty estimates.

Sect. 3.2 will consider the ensemble mean of the Northern Hemisphere reconstruction ensemble (Frank et al., 2010), but we may question the consistency of the single reconstructions with one another. The reconstruction sub-ensemble recalibrated to 1920–1960 is only probabilistically consistent with respect to the recalibrated Frank et al. (2007) reconstruction (Fig. 2). Here, we consider the target uncertainty and account for the reduced internal variability in the data by Hegerl et al. (2007), ~~yet, be sufficient for climate reconstructions on another. Implications and origins of found consistency or lack thereof are discussed~~Mann and Jones (2003) and Mann et al. (2008). The results notably differ, if we exclude the uncertainty estimate (not shown).

Thus, we see from Figure 1a that pairs of ensemble and verification 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 target variability and the projected variability are similar. Figure 2 additionally highlights that the reconstruction ensemble apparently does not generally comply with these assumptions.

3.2 Area-averaged time series

3.2.1 ~~Ensemble consistency of area-averaged estimates~~

~~Figure 3 displays the~~ Figures 3 to 5 display the verification data time series and their variability together with the range of the ensembles. ~~Their probabilistic consistency is illustrated by Fig. 6 and the climatological component of consistency by Fig. 7. The bottom (top) rows of Figs. 6 and 7 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~~ We see that the European data for the simulations and the reconstruction cover a similar range and show similar variability (Fig. 3), while the hemispheric reconstruction ensemble mean varies less than the simulation ensemble and displays a different temporal evolution (Fig. ~~6a), 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 5). Similarly the hemispheric simulation ensemble mean differs in the temporal evolution from the reconstructions but, on the other hand, is in the range of their variability (Fig. 6e). On the other hand, 5).

The range of possible reconstructed evolutions covers a notably wider range than the simulation ensemble estimates for the Northern Hemisphere temperature are from a notably too wide probabilistic distribution relative to the ensemble mean reconstruction if we include the estimates of internal variability in the reconstructions (Fig. 6e5). The bootstrapped ranks (shading in inclusion of an estimate of internal variability does not excessively change the simulation ensemble mean climate trajectory (Fig. 6)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 outcome5). Nevertheless it provides a pronounced increase in the variability of the simulation ensemble mean target. Sections 3.2.2 and 4 discuss the influence of the resolved variability on the results.

Uncertainty estimates for the target data time series The probabilistic consistency is illustrated by Fig. 6 and the climatological component of consistency by Fig. 7. Both figures account for the uncertainty in the verification target. Uncertainty estimates are the reported standard errors for the Central European temperature data target (Dobrovolný et al., 2010) and the spread of the mutual ensembles for the Northern Hemisphere data . Accounting for these "errors" in targets. If we neglect these "observational" uncertainties in the verification data the "verification " data alters the result for the reconstruction ensemble . The ranks in results change for the hemispheric data (not discussed).

3.2.1 Ensemble consistency of area-averaged estimates

Visually, no probabilistic differences arise between the ensemble simulated and reconstructed estimates for the Central European temperature (Fig. 6a). Nevertheless, the χ^2 statistics for the spread-test are significant which would imply a lack of consistency. The bootstrapped rank count intervals however are not incompatible with a uniform result. The contrast between bootstrap and goodness-of-fit test possibly highlights the problem of sampling variability.

Inferences on probabilistic consistency for the hemispheric data depend on whether internal variability is accounted for in the target data with differing results for the reconstruction and simulation ensembles (Fig. 6f clearly display strong over-dispersion,that is, the ensemble mean simulation populates b,c). We first consider the case where the assessment does not include the estimates of internal variability. Then, the reconstruction ensemble mean occupies too often the central ranks of the histogram. This behavior is also found for the ensemble mean reconstruction in and consequently we may term the full simulation ensemble significantly over-dispersive. The bootstrapped intervals confirm this (cyan overlay in Fig. 6b). Under the same conditions also the reconstruction sub-ensemble (calibration period 1920–1960, see Frank et al., 2010) for the northern hemispheric mean temperature is

too wide relative to the ensemble mean simulated Northern Hemisphere temperature (Fig. 6~~d. The bootstrapped ranks and the goodness-of-fit test unambiguously indicate a lack of consistency due to over-dispersive distributions for~~ c).

If we include an estimate of the internal variability in the analyses of the hemispheric data~~.~~

~~No large changes are found in the ranks for the European temperature data (~~, mixed results are obtained. The simulation ensemble may be consistent with some of the constructed reconstruction-targets, but the median of assessments against all targets still emphasises an over-dispersive relation. Furthermore, the 90% envelope (dark grey in Fig. 6b) ~~and the 99range of the bootstrapped ranks is still compatible with a flat histogram. Contrarily,~~ is incompatible with probabilistic consistency. We note that we cannot reject consistency according to the 99% envelope (light grey in Fig. 6b).

On the other hand, if we consider the internal variability for the ~~presented χ^2 test gives significant p-values for spread-deviations, which highlights the problem of sampling variability and the strictness of~~ simulation ensemble mean target in the evaluation of the ~~χ^2 test~~reconstruction ensemble, the analysis suggests that the sub-ensemble recalibrated to the period 1920–1960 is indeed consistent with the simulation ensemble mean (continuous black line in Fig. 6c). The bootstrapped ranks emphasise the good probabilistic agreement under the assumptions made (grey shading in Fig. 6c). Results are insensitive to the inclusion of an arbitrarily chosen estimate of internal variability in the data by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008).

~~Similarly, the~~ The residual quantiles of the climatological distributions in Fig. 7a ~~agree generally well for~~ indicate good agreement between simulated and reconstructed European temperatures~~, although the simulations~~ . Some simulations appear to underestimate very warm ~~annual~~ anomalies and overestimate very cold ~~ones. The time series in Fig. 3a relates the underestimation of the warm anomalies particularly to reconstructed extreme warmth in the mid 16th century. The overestimation of cold anomaliesis more frequent but originates from only few ensemble members (Fig. 7a). If we include the error estimates~~anomalies. Overall, a slight positive 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.~~ , which is indicative of over-dispersion. However, bootstrapped intervals still include the zero line, which clarifies that the slope is not significant.

Larger climatological deviations occur between the simulation ensemble and the reconstructions ~~occur~~ for the Northern Hemisphere temperature data~~(~~. If we do not account for the reduced internal variability in the ensemble reconstruction mean target, the simulation ensemble gives overly wide distributions (grey overlay in Fig. 7b~~, ~~). Similarly, reconstruction ensemble members generally overestimate at least the positive anomaly quantiles relative to the simulation ensemble mean target excluding the internal variability estimate (transparent grey in Fig. 7c). ~~Independent of~~

Results change for the northern hemispheric data under considerations on the ~~reconstruction uncertainty,~~ internal variability. Figure 7b plots residual quantiles for the simulation ensemble ~~gives overly wide distributions. Similarly,~~ relative to 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 uncertaintiesare included~~mean target with added estimates of internal variability for all calculated estimates. In the multitude of possible patterns we can find consistent residuals as well as under-dispersive (negative slope) or over-dispersive ones (positive slope). Furthermore, simulated quantiles appear to commonly agree in the tails but to overestimate the variability closer to the mean. These also include simulated quantiles which besides being more variable close to the mean feature lighter tails. The overestimation appears to be largest for the strong forcing simulations. From our point of view, the multitude of possible deviations leads to a conditional rejection of climatological consistency especially due to the notable overestimation of variability. Rejecting consistency is in line with the probabilistic assessment in Fig. 6b already stressing the generally over-dispersive character of the ensemble.

The analysis of the climatological consistency for the reconstruction sub-ensemble details that most reconstructions agree well with the simulation ensemble mean target when we include an estimate of internal variability in the assessment (Fig. 7c). The bootstrapped intervals emphasise this general consistency. However, deviations are notable in the tails, which become pronounced for large negative anomalies and the reconstruction by D'Arrigo et al. (2006). Residuals for the data by Jones et al. (1998) diverge from the common description by being strongly over-dispersive. The strength of the over-dispersion originates in the size of the considered uncertainties.

~~Considering the two~~ Next, we shortly discuss the two individual full-forcing simulation sub-ensembles ~~separately~~ (five simulations with weak, three with strong solar forcing)~~confirms~~. Respective analyses confirm the results with respect to the European temperature data although both ensembles display specific ~~behaviors~~ behaviours (not shown). ~~If uncertainties in the truth are accounted for, the~~ The weak solar full-forcing ensemble is unambiguously probabilistically consistent with the European reconstructions, whereas the strong solar forcing ensemble is slightly too wide (not shown). 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. 7~~(red, weak ensemble, blue, strong ensemble). Relative~~ b.

The weak solar full-forcing ensemble is likely too wide probabilistically relative to the Northern Hemisphere temperature reconstruction ensemble mean (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~~, including internal variability and considering uncertainty). For the strong solar full forcing ensemble the bootstrapped quantiles and the small ensemble size allow only ambiguous statements although the single deviation test for spread and the rank counts suggest significant over-dispersion (not shown). The residual quantiles display ~~strong~~ a wide range of possible deviations for the strong forcing ensemble (compare Fig. 7~~). Reversing the verification task and considering errors in the truth~~b).

For the reversed verification on hemispheric data, the reconstruction ensemble distribution is ~~too wide~~ slightly (strongly) too narrow relative to the weak ~~forcing ensemble but is consistent relative to the strong forcing ensemble~~ (strong) solar full-forcing ensemble mean according to the χ^2 statistics and the rank histograms if we consider the uncertainties and the internal variability (not shown). However, the bootstrapped quantiles again prevent unambiguous conclusions on the relation between ensemble and verification data. Climatologically most reconstruction ensemble members are consistent with the weak and the strong solar full-forcing ensemble means if uncertainties and internal variability are considered (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~~The results for the reconstruction by Jones et al. (1998) are again distinct from those for the other reconstructions. That is, the climatological deviations relative to the simulation sub-ensembles generally agree with those for the full ensemble displayed in Fig. 7c, but deviations are ~~outside the 99range for the positive tails otherwise.~~ larger relative to the full ensemble mean target.

Next, we shortly give results on the assessment of pairs of simulation ensembles (all, weak, strong solar full-forcing) and single reconstructions (Frank et al., 2010, recalibrated to the 1920–1960 period). We include uncertainty estimates. Furthermore, we choose an arbitrary member of the ensemble of internal variability estimates to add to the three reconstructions by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008). Figure 8 presents the χ^2-values for the tests.

Obviously, the full ensemble lacks probabilistic consistency with all reconstructions under the made assumptions on internal variability and uncertainty, according to the χ^2 test. The bootstrapped intervals confirm this (not shown). Deviations are least obvious for the data by Moberg et al. (2005). Climatological quantiles confirm these probabilistic findings (not shown).

The weak solar full-forcing ensemble appears to be probabilistically consistent with the Moberg et al. (2005) reconstruction. The bootstrapped intervals suggest that the ensemble is not probabilistically inconsistent with the data by Mann et al. (2008) under the made assumptions (not shown). Residual quantiles are generally large (except for Moberg et al., 2005).

The ~~reconstruction quantile residuals relative to the full simulation ensemble mean quantiles (Fig. 7e, f)present an amplified picture of the deviations relative to the two sub-ensembles~~three-member strong solar full-forcing ensemble is a special case. Bootstrapped intervals do not permit to reject probabilistic consistency for any of the nine reconstructions under the assumptions made. Results summarized in Figure 8c indicate consistency of the ensemble with the data by Frank et al. (2007), Moberg et al. (2005) and Mann et al. (2008). Again, residual quantiles are large except for the reconstruction by Moberg et al. (2005). We note that the results differ for all three ensembles (all, weak, strong solar full-forcing) if we do not include uncertainty estimates.

If the surrogate ensemble generated from control-run data (see Sect. 3.1) is assessed against the

521 members of the Frank et al. (2010, uncertainty estimate included) recalibration ensemble, about 14 % of the pairs arise as consistent with respect to the full test although they are unrelated (Figure 1b). Single spread test statistics are not significant in about 35 cases (Figure 1b). We include an estimate of internal variability for the reconstructions by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008). Similarly, the climatological analyses displays larger consistency for the surrogate ensemble than for the real ensemble, with some members of the reconstruction ensemble (not shown) indicating strong deviations between reconstructed and simulated climate evolutions. Thus, the unperturbed internal climate variability may be indistinguishable from forced simulated or reconstructed variability.

~~Thus~~In summary, 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~~ likely too wide relative to the northern hemispheric mean temperature reconstructions. 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~~ There, the ~~Central European temperature estimates. On the other hand~~results are very diverse relative to the possible representations of internal variability for the reconstruction ensemble mean target. When we account for uncertainties and internal variability, 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~~appears to be consistent with the simulation ensemble mean target but most reconstruction ensemble members deviate climatologically in the tails. Thus, the large uncertainties in the ensembles and also in the verification targets prohibit to generally reject consistency for the northern hemispheric data. On the other hand, the Central European temperature estimates appear to be unambiguously consistent.

3.2.2 Addressing origins of the lack of consistency

~~Figure 3 displays (i) that~~ As described above, the European data ~~for the simulations and the reconstruction~~ cover a similar range and show similar variability ~~, (ii) that the hemispheric reconstructionensemble mean varies less than the simulation ensemble and displays different temporal evolution, as does (iii) the hemispheric simulation ensemble mean (~~for the simulations and the reconstruction. Figure 3b further displays that the low-frequency variability differs notably between the simulations and the reconstruction.

Figure 5 and 5 also highlight prominent differences between the hemispheric targets and the hemispheric ensemble data. When we account for the reduced internal variability in the hemispheric

ensemble mean targets we find that the range of possible reconstructed evolutions is relatively wide compared to the ~~reconstruction ensemble), which on~~ simulation ensemble. The moving standard deviations emphasise the disagreement in variability. On the other hand~~is in the range of~~ , the inclusion of an estimate of internal variability does not unduely change the simulation ensemble mean climate trajectory, but it provides a distinct increase in the variability of the ~~reconstruction ensemble . However,~~ simulation ensemble mean target (Fig. 5b).

We note that 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 . 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. 3c 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~~ externally-forced climate variability for the continent.

On the other hand, the ~~distributions differ between the northern hemispheric~~ probabilistic and climatological evaluations emphasise the disagreement between the Northern Hemisphere temperature reconstruction ensemble mean and the full simulation ensemble~~, if we consider the uncertainty in the verification ensemble mean reconstruction~~. The time series in Figures 5 and 5 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.~~ but are not large enough to reject consistency. They rather compensate over the full period.

~~Figure 3 further shows that the considered ensembles of estimated temperature anomaly series generally enclose the verification data (Fig. 3a–c), but they often over-estimate inter-annual variability (Fig. 3d–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. 3 ~~a, b~~and 5). 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. 3e, 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~~

Appropriate representation of internal variability is fundamental for our assessment, and internal climate variability can be as large as forced signals. In Figures 6 and 7, we followed different approaches. As mentioned before, panels b–c of both figures display the assessment of the respective ensemble relative to an ensemble-mean estimate which presents only a reduced amount of internal variability. Even after including the estimates of the internal variability, results for simulations and reconstructions describe to some extent different aspects of our uncertain knowledge. While the spread of the ~~target simulation ensemble mean (Fig. 3f). 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. 3e, f, e.g. 13th, 15th and early 19th centuries, compare also , and~~ reconstruction ensemble relates to different methodologies and different climate proxies, the intra-ensemble variability for the simulations represents the differences in the considered forcing estimates and the different initial conditions of the ensemble but also depends on the formulation of the numerical code. The added internal variability for the simulation target describes one unperturbed climate trajectory under similar constraints. The internal variability adjustments for the reconstructions may still represent the different methodologies and types of proxy data although they are generated as stochastic processes.

In our analyses, we accounted for the reduced internal variability in ensemble mean targets. However, we note that strong discrepancies in variability may also occur with respect to inter-annually representative temperature reconstructions (not shown).

3.3 Spatial fields

3.3.1 ~~Ensemble consistency of field estimates~~

In the following, we extend the analyses of consistency ~~are extended~~ to the decadally smoothed global temperature ~~field reconstruction by . We~~ fields. We thus 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. 9 a, and~~ Sub-periods of non-overlapping 250 records are considered in the range from 805 to 1845 CE. The first three periods cover the first 750 records of the full data (about 800 to 1050, 1050 to 1300, 1300 to 1550), but the last period covers the last 250 records of the data sets (about 1595 to 1845). Thus, there is a gap between the earlier three periods and the late period.

Figure 9 provides a first impression of the relation between simulated and reconstructed data for the global temperatures. Fig. 9c shows the rank of the reconstruction data in the COSMOS-Mill ensemble of surface temperature data for this climatology. From a displays the reconstructed climatology map for an arbitrarily chosen sub-period (1390s to 1690s) from the decadally smoothed global temperature field data (reconstruction by Mann et al., 2009). The ranks in Fig. 9c suggest strong deviations between the ensemble and the reconstruction over wide regions of the globe for this sub-period with the ensemble being biased low over the tropical Pacific ocean and high over most other oceanic regions, North America and eastern and western Eurasia. These biases are not representative as we discuss below (compare Fig. 12). Rather the ranks in Fig. 9c highlight how strongly simulated mean anomalies may disagree with the reconstructed patterns for specific periods. Variability is as often comparable as not for the data set not only in the sub-period but also over the full period (Fig. 9bit can be seen that the simulations frequently vary more than the field reconstruction at individual grid points.). Sub-periods generally give similar patterns of relative standard deviations. However, slight changes may of course be found in the specific size of over- or under-estimation of variations in the sub-periods.

3.3.1 Ensemble consistency of field estimates

Figures 10 to 12 display a selection of results for the evaluation of consistency . Although no at individual grid-points. No uncertainty estimate is given for the global field data, we so we choose to inflate the ensemble by a random error estimate drawn from a distribution with a standard deviation equaling standard deviation equal to the largest standard error of the unscreened Northern Hemisphere mean temperature series provided by Mann et al. (2009). Without error uncertainty inflation, expected effective rank frequencies can be very small considering due to 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. 10 for a random selection of grid points) and residual quantiles (Fig. 11 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 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 1845CE sub-periods (Fig. 10). 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. 11). Among the individual ensemble members,

~~the climatological behavior~~ The climatological behaviour is mostly comparable among the individual ensemble members 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~~ between the sub-periods are diverse but can be rather small between the first and the last 250 records (compare Fig. 11). Some improvement is seen towards more limited deviations or nearly vanishing residuals in the ~~late period~~last sub-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~~the first and the last sub-period. Thus, results for sub-periods are often not comparable with each other in ~~neither the probabilistic nor the climatogic~~ either the probabilistic or the climatological evaluation. Occurring ~~shifts emphasize~~ subsequent shifts emphasise the general lack of a common signal, i.e. differences in the long-term trend component.

Decadal smoothing reduces the width of the climatological quantile distributions, and a number of grid points display only ~~very small~~ a very small range of quantiles as a sign of very weak inter-decadal variability (not shown). ~~At certain grid points, the~~ The extremely narrow reconstructed quantile distributions result in particularly strong climatological over-dispersion at certain grid points. Quantile distributions are in parts broader in higher Northern Hemisphere latitudes for both 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~~ selection of grid-points provides only a snapshot of the results for the global field data. Fig. 12 ~~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~~ provides a summary of the full and single deviation goodness-of-fit tests for the full period and the ~~truth are considered (not shown)~~sub-periods defined above. We include the target uncertainties in all results displayed, but first discuss the results without them. Then, 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~~sub-periods as defined above. Thus, while ~~centering~~ centring 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~~ differ notably at the beginning and at the end.

If a moderate random ~~error~~ uncertainty inflation is used ($\sigma = 0.1729$, see Sect. 2.2), spatially extended ~~consistency~~ probabilistic consistency for the full period is mainly restricted, according to

the full test, to Central Eurasia and the Tropical Pacific ~~for the full period~~ (Fig. 12a). ~~For four~~ Diverging results become visible in the 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 (~~sub-period (about 800 to 1050, Fig. 12b), but uniformity is rejected for the following ~~250 records~~ one (Fig. 12c). 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~~ sub-period from about 1300 to 1550 (the early Little Ice Age, Fig. 12d), the ensemble appears to be consistent in Northern North America, the Tropical Pacific and South of Greenland. In the last period (Fig. 12e, 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~~target. On the other hand, single deviations are nearly always and everywhere significant (Fig. 12f–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 (ranks plotted in Fig. 9d) if we consider the uncertainties in the reconstruction. Otherwise the rank histogram displays an overpopulation of central and outer ranks, which is an effect of the accumulation of the individual grid point deviations (not shown). The ~~cumulative spatial assessment suggests~~ different biased, under- and over-dispersive relations suggest strongly differing relations between reconstructed and simulated decadal temperatures ~~on global scales (not shown)~~in different regions.

In summary, ~~as for the~~ even more prominent than for the area-averaged 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~~ Lack of consistency is ~~seen~~ most prominent over the southern oceans. Tests for the single deviations of bias and spread are nearly everywhere significant after inclusion of an ~~error estimate~~ uncertainty estimate following our description in Sect. 2.2. 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.

3.3.2 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. 9~~c which refer to a~~ a presents the reconstructed mean anomaly map for an arbitrary sub-period ~~of the Little Ice Age~~ (1390s to 1690s) ~~. The reconstructed climatology map for this period is shown~~ encompassing part of the Little Ice Age. Mapped ranks in

Fig. 9~~a. While the~~ c exemplify possible differences in simulated and reconstructed mean anomalies patterns. The amplitudes of mean anomalies are comparable between reconstructions and strong solar full-forcing simulations except in the Tropical Pacific, but the weak solar full-forcing simulations display less cooling in the selected period (not shown, compare example map in Fig. 9a and rank map in Fig. 9c). ~~Variability is as often comparableas not (Fig. 9b). The simulations especially vary more~~ While variability is often comparable, the simulations display more variability than the reconstruction over oceanic regions (~~middle blue in~~ see Fig. 9b). 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. 9b.

The ranks in Fig. 9c 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~~ sub-periods as exemplified by the decadally smoothed global data over the ocean regions (see mapped ranks in Fig. 9c). This holds also for other field reconstructions (not shown). ~~It is most pronounced over the oceans for the decadally smoothed global data (Fig. 9c).~~ 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~~ is 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~~ reduces the subjectivity of a comparison between simulations and reconstructions, and the goodness-of-fit test allows to ~~summarise, in one Figure,~~ summarize the (dis-)agreement in terms of ~~ensemble consistency.~~ ensemble consistency in one figure.

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~~ . They also highlight periods in which the simulations are rather warm compared to temperature reconstructions when ~~anomalies~~ temperature deviations 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~~ we realign the simulations and the reconstructions to the mean of a common period to correct systematic differences in 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 .~~ trends before applying tests of consistency (similarly 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). We accept that the choice of the reference period influences the results.

~~*Further data sets:* strong~~ Strong probabilistic and climatological deviations can arise between the ~~data presented above~~ simulations and the reconstructions 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~~ also indicate a generally over-dispersive character of the ensemble (not shown). ~~We can generally~~ Even if we cannot reject uniformity at the grid point level and for area average series the associated uncertainties lessen the value of such consistency. 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 , 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 , and 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 , the simulation ensembles display least deviations relative to the data by and . The three-member strong solar full-forcing ensemble appears also to be consistent with the , ,~~ fully consistent between the simulation ensemble and ~~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 . In line with similar considerations in seasonal and medium-range weather forecasting , 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 recalibration ensemble, about 20of 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 reconstructionsub-ensemble calibrated to 1920–1960, indicating strong deviations between forced reconstructed and simulated climate evolutions~~the reconstruction.

~~*Further discussions:*~~ Thus, 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~~ used to drive the simulations can also be assumed to be less uncertain in this period 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 and the related teleconnections 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~~representing inter-annual variations. 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~~ annual variability. Results change notably whether uncertainties and/or internal variability estimates are included in the assessment of the reconstruction sub-ensemble against the specific simulation ensemble mean~~with this particular reconstruction ensemble , but the larger variability in the simulations compromises the inverse consideration~~. Although the temporal evolutions notably deviate, it appears likely that the reconstruction ensemble and most of its members are indeed consistent, i.e. comparable, with the chosen ensemble simulation mean verification target under the assumptions made on internal variability and the uncertainties. However, the simulation ensemble displays pronounced deviations from consistency relative to the ensemble mean reconstruction target including various estimates of internal variability. Interestingly, the moving standard deviations of the ensemble means (simulations and reconstructions) evolve to some extent similarly in the period 1400 to ~~1900.~~ 1900 (compare Figs. 3–5). The 20th century disagreement is possibly due to the evolution of the simulations with strong solar forcing. Including estimates for internal variability introduces an additional source of uncertainty. While it reduces the problems in employing ensemble mean targets, it also highlights the ambiguity of our estimates of past climate trajectories.

~~With a focus similar to the approach utilized here,~~ Sundberg et al. (2012) and 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~~ goal is similar to the approach utilized here. Their framework focusses 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~~ assess 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. We propose that, in parallel, we need to address the compatibility of reconstructions and simulations by evaluating their probabilistic and climatological consistency.

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 for 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~~study.

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~~ a verification data set (e.g. Annan and Hargreaves, 2010; Marzban et al., 2010). If no ~~state of truth~~ reliable observable target can be identified, as is the case in periods and regions without instrumental observations, such statistical analyses ~~add an objective component to~~ reduce the subjectivity of the evaluation of simulation ensembles and statistical approximations from paleo-sensor data (Braconnot et al., 2012) under uncertainty and go beyond "wiggle matching". The approach permits a succinct visualization of the consistency between an ensemble of estimates and an uncertain verification ~~truth~~target. 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~~ We consider the COSMOS-Mill-ensemble (Jungclaus et al., 2010) and various reconstructions within the described approach~~, we~~ . 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~~ The ensemble possibly 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~~ over-dispersion and various climatological deviations. The ensemble generally samples from a significantly wider distribution than the reconstruction ensemble mean. The distribution of the reconstruction ensemble in turn is ~~too wide~~ possibly consistent relative to the simulation ensemble mean.

~~Similarly~~Furthermore, 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~~ the data is probabilistically consistent for multi-centennial sub-periods and certain regions according to the applied full test, ~~accounting for~~ analyses of single probabilistic deviations and climatological differences ~~emphasizes~~ emphasise a general lack of consistency. ~~The~~ We find the largest, but still limited~~consistency is seen~~ , consistency 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~~true past climate state and evolution.

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~~ which 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. Two anonymous referees helped to strengthen our manuscript and comments by T.L. Edwards also lead to notable improvements.

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 it contributes to the Cluster of Excellence "CliSAP" at ~~Hamburg University~~ the University of Hamburg.

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.

Annan, J. D. and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, http://dx.doi.org/10.1029/2009GL041994doi:10.1029/2009GL041994, 2010.

Annan, J. D., Hargreaves, J. C., and Tachiiri, K.: On the observational assessment of climate model performance, Geophys. Res. Lett., 38, L24702, http://dx.doi.org/10.1029/2011GL049812doi:10.1029/2011GL049812, 2011.

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., Lan, 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, http://dx.doi.org/10.5194/cp-3-261-2007doi:10.5194/cp-3-261-2007, 2007.

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, http://dx.doi.org/10.1038/nclimate1456doi:10.1038/nclimate1456, 2012.

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, http://dx.doi.org/10.1037/0033-2909.86.6.1290doi:10.1037/0033-2909.86.6.1290, 1979.

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, http://dx.doi.org/10.1007/978-1-4020-5725-0_1doi:10.1007/978-1-4020-5725-0_1, 2011.

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, http://dx.doi.org/10.1007/s10584-009-9783-zdoi:10.1007/s10584-009-9783-z, 2010.

Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M. and Bladé, I.: The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field, J. Climate, 12, 1990–2009, 1999.

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

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, http://dx.doi.org/10.1038/30943doi:10.1038/30943, 1998.

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, http://dx.doi.org/10.5194/essdd-5-1-2012doi:10.5194/essdd-5-1-2012, 2012.

D'Arrigo, R., Wilson, R., and Jacoby, G.: On the long-term context for late twentieth century warming, J. Geophys. Res., 111, D03103, http://dx.doi.org/10.1029/2005JD006352doi:10.1029/2005JD006352, 2006.

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, http://dx.doi.org/10.1007/s10584-009-9724-xdoi:10.1007/s10584-009-9724-x, 2010.

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.

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, http://dx.doi.org/10.1029/2006PA001304doi:10.1029/2006PA001304, 2007.

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, http://dx.doi.org/10.1029/2007GL030571doi:10.1029/2007GL030571, 2007.

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, http://dx.doi.org/10.1038/nature08769doi:10.1038/nature08769, 2010.

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, http://dx.doi.org/10.1029/2008JD010239doi:10.1029/2008JD010239, 2008.

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

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, http://dx.doi.org/10.5194/cp-7-917-2011doi:10.5194/cp-7-917-2011, 2011.

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, http://dx.doi.org/10.1175/JCLI4011.1doi:10.1175/JCLI4011.1, 2007.

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. PastDiscuss.~~ Part 2: A pseudo-proxy study addressing the amplitude of solar forcing, Climate of the Past, 8, ~~263–320,~~ 1355–1365, http://dx.doi.org/10.5194/cp-8-1355-2012doi:10.5194/cp-8-1355-2012 doi:10.5194/cp-8-1355-2012, 2012.

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.

Johnson, C. and Bowler, N.: On the reliability and calibration of ensemble forecasts, Mon. Weather Rev., 137, 1717–1720, http://dx.doi.org/10.1175/2009MWR2715.1doi:10.1175/2009MWR2715.1, 2009.

Jolliffe, I. T. and Primo, C.: Evaluating rank histograms using decompositions of the Chi-square test statistic, Mon. Weather Rev., 136, 2133–2139, http://dx.doi.org/10.1175/2007MWR2219.1doi:10.1175/2007MWR2219.1, 2008.

Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B.: High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with General Circulation Model control-run temperatures, The Holocene, 8, 455–471, doi:10.1191/095968398667194956, 1998.

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.

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, http://dx.doi.org/10.5194/cp-3-591-2007doi:10.5194/cp-3-591-2007, 2007.

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, http://dx.doi.org/10.1175/JCLI3827.1doi:10.1175/JCLI3827.1, 2006.

Jungclaus, J. H., Lorenz, S. J., Timmreck, C., Reick, C. H., Brovkin, V., Six, K., Segschneider, J., Giorgetta, M. A., Crowley, T. J., Pongratz, J., Krivova, N. A., Vieira, L. E., Solanki, S. K., Klocke, D., Botzet, M., Esch, M., Gayler, V., Haak, H., Raddatz, T. J., Roeckner, E., Schnur, R., Widmann, H., Claussen, M., Stevens, B., and Marotzke, J.: Climate and carbon-cycle variability over the last millennium, Clim. Past, 6, 723–737, http://dx.doi.org/10.5194/cp-6-723-2010doi:10.5194/cp-6-723-2010, 2010.

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, http://dx.doi.org/10.1007/s00382-001-0196-6doi:10.1007/s00382-001-0196-6, 2002.

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, http://dx.doi.org/10.1126/science.1093877doi:10.1126/science.1093877, 2004.

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, http://dx.doi.org/10.1007/s10584-009-9782-0doi:10.1007/s10584-009-9782-0, 2010.

Mann, M. E., and Jones, P. D.: Global surface temperatures over the past two millennia, Geophysical Research Letters, 30, 1820, doi:10.1029/2003GL017814, 2003.

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, http://dx.doi.org/10.1073/pnas.0805721105doi:10.1073/pnas.0805721105, 2008.

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, http://dx.doi.org/10.1126/science.1177303doi:10.1126/science.1177303, 2009.

Mann, M. E., Fuentes, J. D., and Rutherford, S.: Underestimation of volcanic cooling in tree-ring-based reconstructions of hemispheric temperatures, Nat. Geosci., 5, 202–205, http://dx.doi.org/10.1038/ngeo1394doi:10.1038/ngeo1394, 2012.

Marzban, C., Wang, R., Kong, F., and Leyton, S.: On the effect of correlations on rank histograms: reliability of temperature and wind speed forecasts from finescale ensemble reforecasts, Mon. Weather Rev., 139, 295–310, http://dx.doi.org/10.1175/2010MWR3129.1doi:10.1175/2010MWR3129.1, 2010.

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, http://dx.doi.org/10.1126/science.1172872doi:10.1126/science.1172872, 2009.

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, http://dx.doi.org/10.1175/JCLI-D-11-00261.1doi:10.1175/JCLI-D-11-00261.1, 2012.

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, http://dx.doi.org/10.1038/nature03265doi:10.1038/nature03265, 2005.

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

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, http://dx.doi.org/10.1007/s00382-010-0793-3doi:10.1007/s00382-010-0793-3, 2011.

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

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.

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, http://dx.doi.org/10.5194/gmd-4-33-2011doi:10.5194/gmd-4-33-2011, 2011.

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, http://dx.doi.org/10.1029/2011GL046658doi:10.1029/2011GL046658, 2011.

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, http://dx.doi.org/10.1051/0004-6361/201016173doi:10.1051/0004-6361/201016173,

2011.

Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, Geophys. Res. Lett., 36, L19704, http://dx.doi.org/10.1029/2009GL040142doi:10.1029/2009GL040142, 2009.

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, Climate of the Past, 8, 1339–1353, http://dx.doi.org/10.5194/cp-8-1339-2012doi:10.5194/cp-8-1339-2012, 2012.

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

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.

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

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, http://dx.doi.org/10.1029/2006JD008318doi:10.1029/2006JD008318, 2007.

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, http://dx.doi.org/10.1029/2005GL023424doi:10.1029/2005GL023424, 2005.

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, http://dx.doi.org/10.1029/2010GL043717doi:10.1029/2010GL043717, 2010.

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, http://dx.doi.org/10.1007/s00382-012-1361-9doi:10.1007/s00382-012-1361-9, 2012.

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, http://dx.doi.org/10.1007/s10584-010-9824-7doi:10.1007/s10584-010-9824-7, 2010.

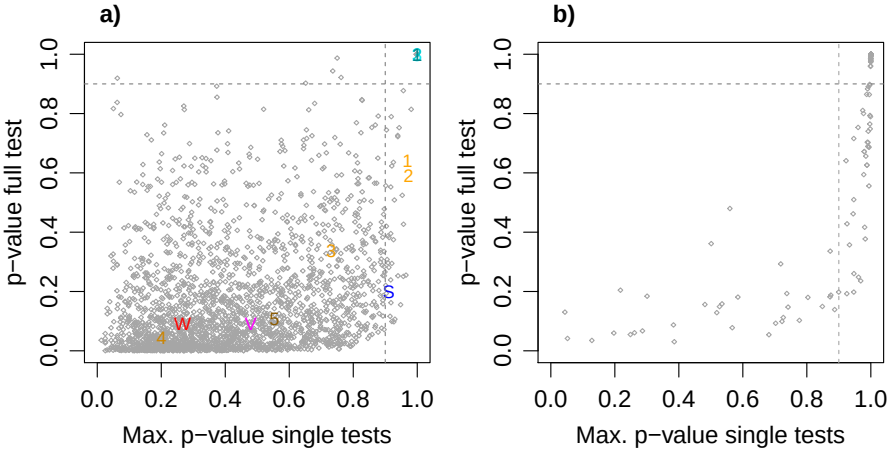


Fig. 1. ~~(a–c)~~ Surrogate ensemble: **(a)** Testing against 2201 surrogate targets: p-values for the full goodness-of-fit χ^2 test plotted against the maximum of the p-values obtained from the tests for bias and spread deviations. In **(a)** orange (blue, see top right corner) numbers 1–5 (1–3) give values for the five full-forcing simulations with weak (strong) solar forcing; red *W*, blue *S* and magenta *V* show values for weak and strong solar forcing only and volcanic only simulations. No uncertainties are considered in **(a)**. **(b)** as **(a)** but for the test against the 521 members of the Frank et al. (2010) ensemble as targets. Horizontal and vertical lines indicate a conservative 10% level for the probability of the rank histograms to agree with the null hypothesis. **(b)** accounts for uncertainties and reduced internal variability in data by Hegerl et al. (2007), Mann and Jones (2003) and Mann et al. (2008).

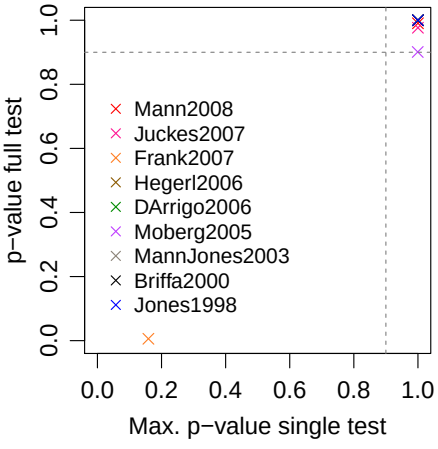


Fig. 2. The Northern Hemisphere reconstruction sub-ensemble re-calibrated to 1920–1960: Test for consistency of the remaining members of the ensemble against a target defined by one of the members: p-values for the full goodness-of-fit χ^2 test plotted against the maximum of the p-values obtained from the tests for bias and spread deviations. Note that the results cluster in the top-right corner of the panel for five of the possible targets. Uncertainty inflation was chosen to be proportional to the full ensemble spread. Results change if we consider only sub-ensemble spread but conclusions remain valid.

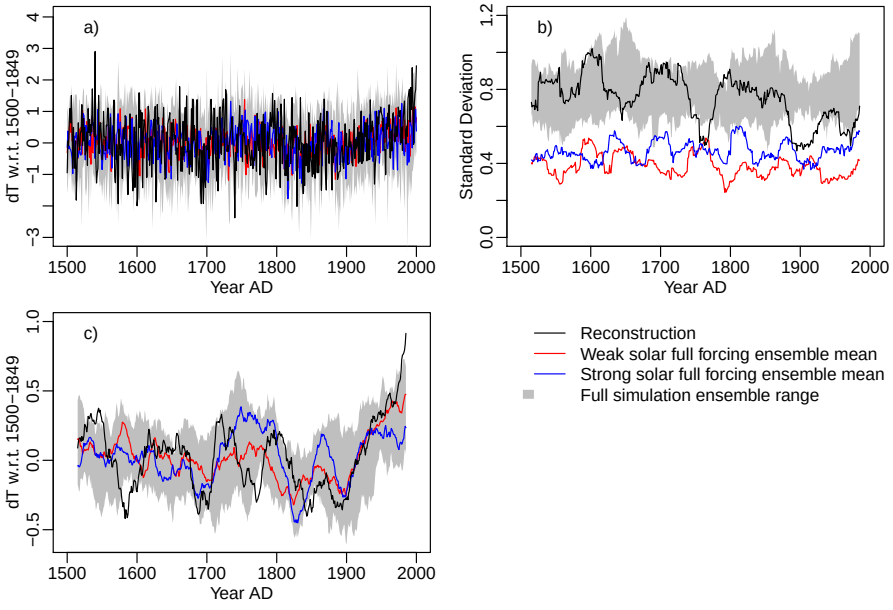
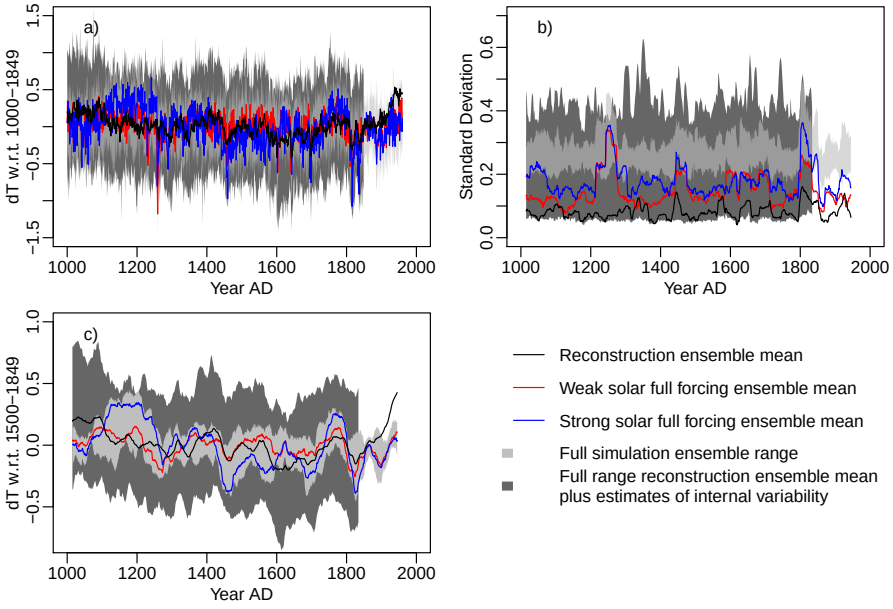


Fig. 3. **(a)** Time series, **(b)** moving 31-yr standard deviations and **(c)** moving 31-yr means for the Central European annual temperature data. Black is the verification data and transparent light grey shading is the range of the ensemble. Red (blue) lines are for the weak (strong) solar full-forcing simulation ensemble means.



~~Rank histogram counts for temperature data including **(a, b)** Central European annual temperatures, **(c, d)** Northern Hemisphere simulation ensemble temperature~~ 5, ~~**(e, f)** Northern Hemisphere reconstruction sub-ensemble calibrated to the period 1920 to 1960. Top (bottom) row does (does not) neglect~~ the ~~errors in~~ data including the ~~truth. Numbers are χ^2 statistics. χ^2 statistics in brackets account~~ estimate for ~~auto-correlation in the data. Grey shading (line) are 0.5and 99.5(50) quantiles~~ internal variability is only shown for ~~block-bootstrapped rank histograms (2000 replicates, block length~~ the period of ~~50 yr). Blue horizontal lines give~~ analysis from the ~~expected average count for a perfectly uniform histogram~~start of the millennium to the mid-19th century.

Fig. 4. **(a)** Time series~~. **(d–f)** Moving~~, **(b)** moving 31-yr standard deviations ~~. **(a, d)** European annual temperature, **(b, c)**~~ and **(c)** moving 31-yr means for the Northern Hemisphere temperature simulation ensemble against the reconstructed ~~truth, **(c, f)** Northern Hemisphere reconstruction ensemble versus simulated truth~~target. ~~In all panels, black~~ Black is the ~~respective~~ verification data and transparent light grey shading is the range of the ~~ensembles~~ensemble. ~~In **(a, b, d, e, f)** red~~ Red (blue) lines are for the weak (strong) solar full-forcing simulation ensemble means. ~~In **(c, f)**~~ Dark grey shading is 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~~ ensemble-mean target with added internal variability estimates~~, and we choose to present the truth~~. Here and ~~the strong solar full-forcing simulation ensemble means~~ in ~~**(a, b)** by dashed lines to increase~~ the ~~visibility of all time series~~Fig.

~~Rank histogram counts for temperature data including **(a, b)** Central European annual temperatures, **(c, d)** Northern Hemisphere simulation ensemble temperature~~ 5, ~~**(e, f)** Northern Hemisphere reconstruction sub-ensemble calibrated to the period 1920 to 1960. Top (bottom) row does (does not) neglect~~ the ~~errors in~~ data including the ~~truth. Numbers are χ^2 statistics. χ^2 statistics in brackets account~~ estimate for ~~auto-correlation in the data. Grey shading (line) are 0.5and 99.5(50) quantiles~~ internal variability is only shown for ~~block-bootstrapped rank histograms (2000 replicates, block length~~ the period of ~~50 yr). Blue horizontal lines give~~ analysis from the ~~expected average count for a perfectly uniform histogram~~start of the millennium to the mid-19th century.

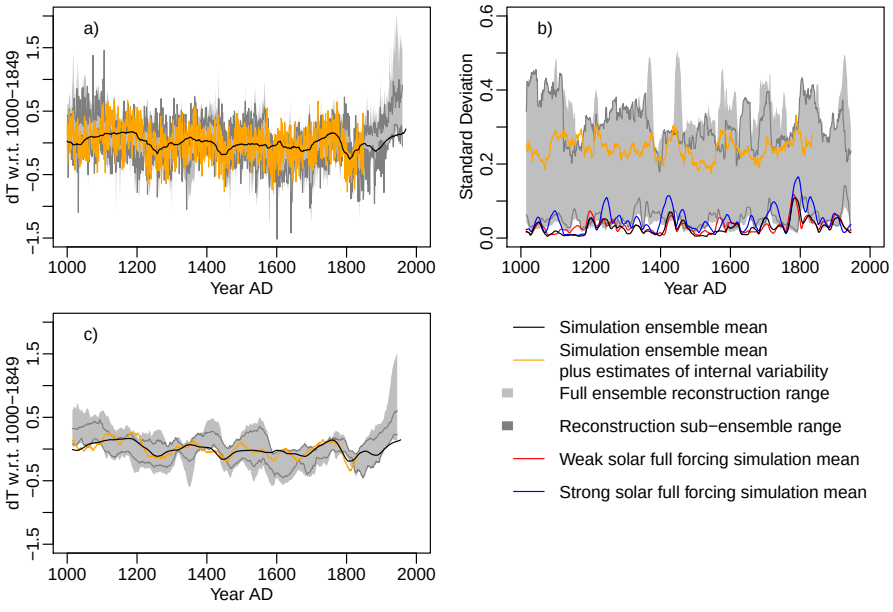


Fig. 5. **(a)** Time series, **(b)** moving 31-yr standard deviations and **(c)** moving 31-yr means for the Northern Hemisphere temperature reconstruction ensemble against the simulated target. Black is the verification data and transparent light grey shading is the range of the ensemble. Dark grey lines mark the range of the reconstruction sub-ensemble recalibrated to the period 1920–1960. The orange line is the estimate of the ensemble-mean target with added internal variability estimate. In **(b)** red (blue) lines are for the weak (strong) solar full-forcing simulation ensemble means.

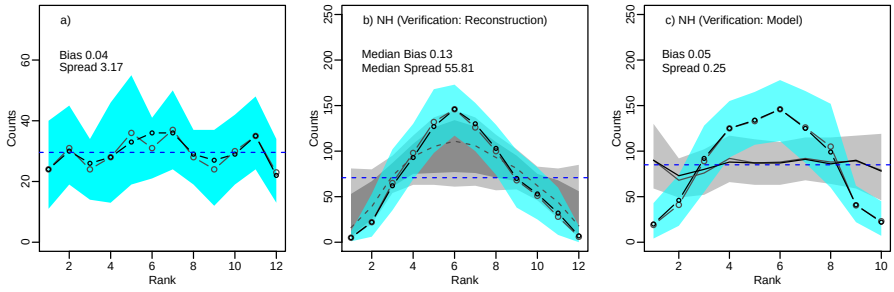


Fig. 6. Rank histogram counts (black line with points) for temperature data: **(a)** Central European annual temperatures, **(b)** Northern Hemisphere simulation ensemble temperature, **(c)** Northern Hemisphere reconstruction sub-ensemble calibrated to the period 1920 to 1960. Analysis does include the uncertainties in the target. Numbers are χ^2 statistics accounting for auto-correlation. In **(b)** they are the median relative to all representations of ensemble-mean reconstruction plus internal variability estimate; in **(c)** they are relative to the ensemble-mean simulation plus internal variability estimate. We use the same representation for equal approaches. Different representations are for different approaches. Cyan shading (grey lines with points) are 0.5 % and 99.5 % (50 %) quantiles for block-bootstrapped rank histograms (2000 replicates, block length of 50 yr). Light grey shading and dashed line in **(b)** are equivalent quantiles for the various estimates of internal variability, dark grey shading adds 5 % and 95 % quantiles. In **(c)** black continuous line is rank count for the ensemble-mean simulation with added internal variability estimate. Grey shading and continuous line add bootstrapped 0.5 % and 99.5 % and 50 % quantiles. Blue horizontal lines give the expected average count for a perfectly uniform histogram. Single test critical values are 2.706 for a χ^2 distribution with one degree of freedom (see Sect. 2.1 and Jolliffe and Primo, 2008; Annan and Hargreaves, 2010) and a conservative one-sided 90% level.

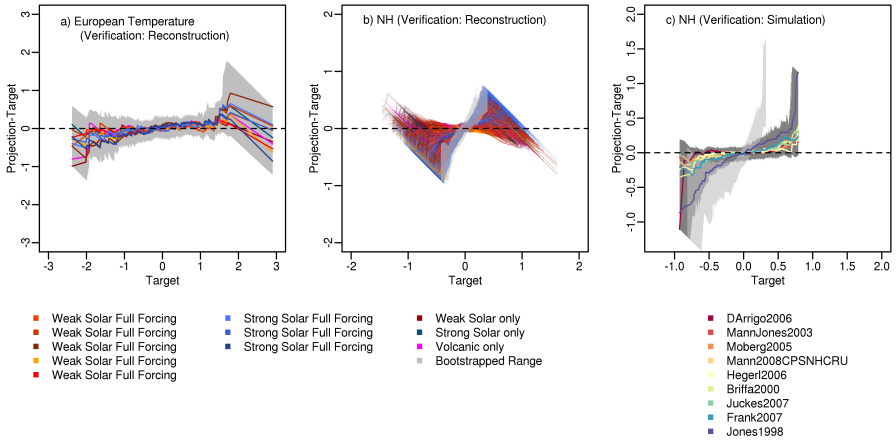


Fig. 7. Residual quantile-quantile plots for temperature data~~including **(a, b)**~~ : **(a)** Central European annual temperatures, ~~**(c, d)**~~ **(b)** Northern Hemisphere simulation ensemble temperature, ~~**(e, f)**~~ **(c)** Northern Hemisphere reconstruction sub-ensemble calibrated to the period ~~1920 to 1960. Top (bottom) row does (does not) neglect~~ 1920–1960. Panels account for the ~~errors~~ uncertainties in the ~~truth~~target. See legend for individual ensemble members. Grey shading in **(a)** and transparent grey overlay in **(b–c)** are 0.5 % and 99.5 % quantiles for block-bootstrapped residual quantiles (2000 replicates, block length of 50 yr). In **(b)** we plot all results relative to all used targets including an estimate of internal variability. In **(c)** the dark grey shading are the bootstrapped quantiles relative to the target including an estimate of simulated internal variability. Middle grey **(c)** is due to the transparency.

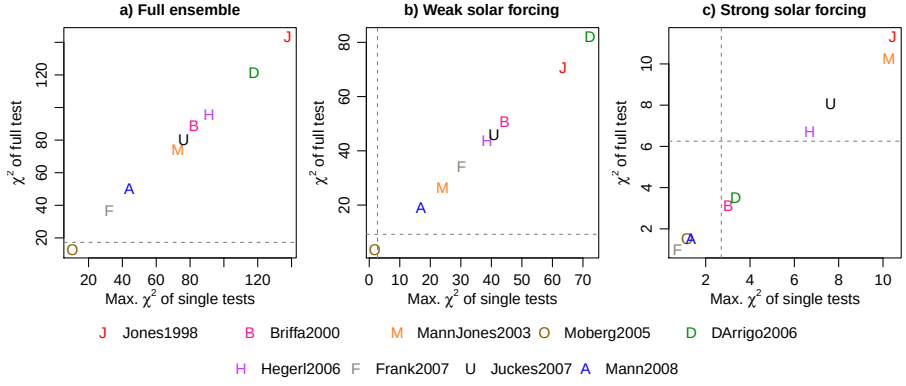


Fig. 8. ~~Global decadal smooth~~ Assessing the full-forcing simulation ensemble against individual reconstructions of Northern Hemisphere temperature~~: **(a)** reconstructed mean anomaly map for a cold period (for~~. Uncertainties are considered, and internal variability estimates are included in the ~~1390s to 1690s)~~data by Hegerl et al. (2007), ~~**(b)** Mann and Jones (2003)~~ and Mann et al. (2008) to account for the temporal filtering of the individual reconstructions. **(a)** Full eleven-member ensemble~~mean~~: χ^2 statistics for the full test against the maximum of ~~relative standard~~ the decomposed χ^2 statistics obtained for the tests for bias and spread deviations~~(reconstruction standard deviation divided by simulation standard deviation at each grid point~~. **(b)** as **(a)** but for the ~~full period),~~ five-member weak solar full-forcing ensemble. **(c)** ~~mapped ranks~~ as **(a)** but for three-member strong solar full-forcing ensemble. Vertical and horizontal grey lines mark those χ^2 statistics for which left p-values are larger than 0.9 for the ~~cold period (1390s to 1690s)~~distributional degrees of freedom.

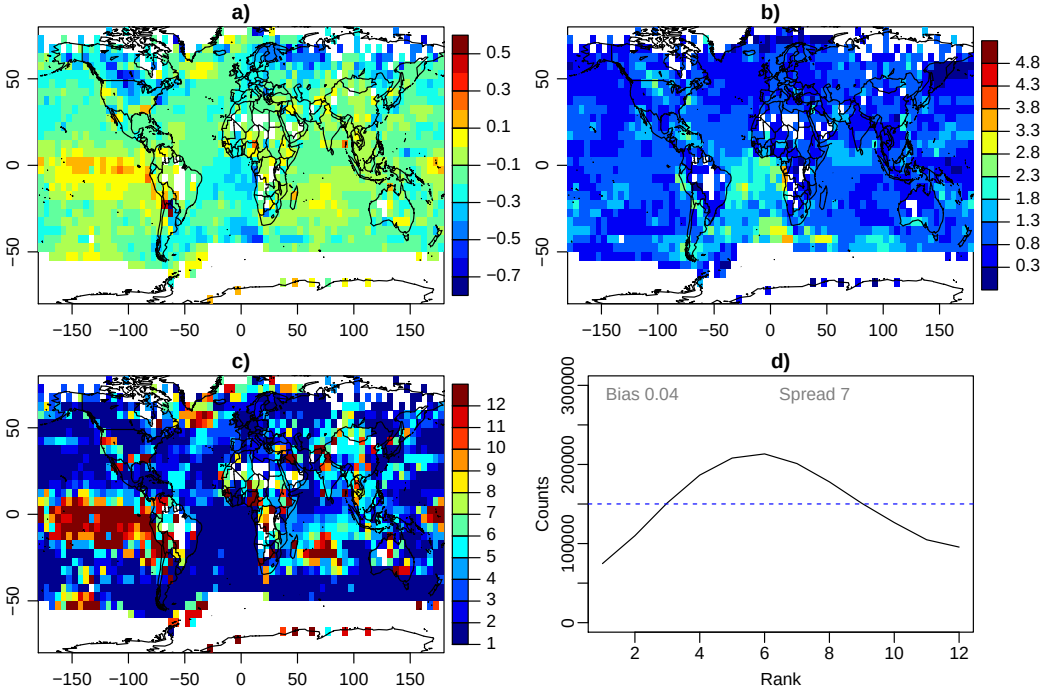


Fig. 9. Global fields of decadally smoothed temperature: **(a)** reconstructed mean anomaly map for a cold period (for the 1390s to 1690s), **(b)** ensemble mean of relative standard deviations (reconstruction standard deviation divided by simulation standard deviation at each grid point for the full period), **(c)** mapped ranks for the cold period (1390s to 1690s). Panel **(d)** presents rank counts for all points in space and time considering uncertainties in the data for the full period. Numbers in **(d)** are χ^2 statistics for the full period accounting for spatial and temporal auto-correlation in the data. We assume 15 spatial degrees of freedom (calculated following Bretherton et al., 1999).

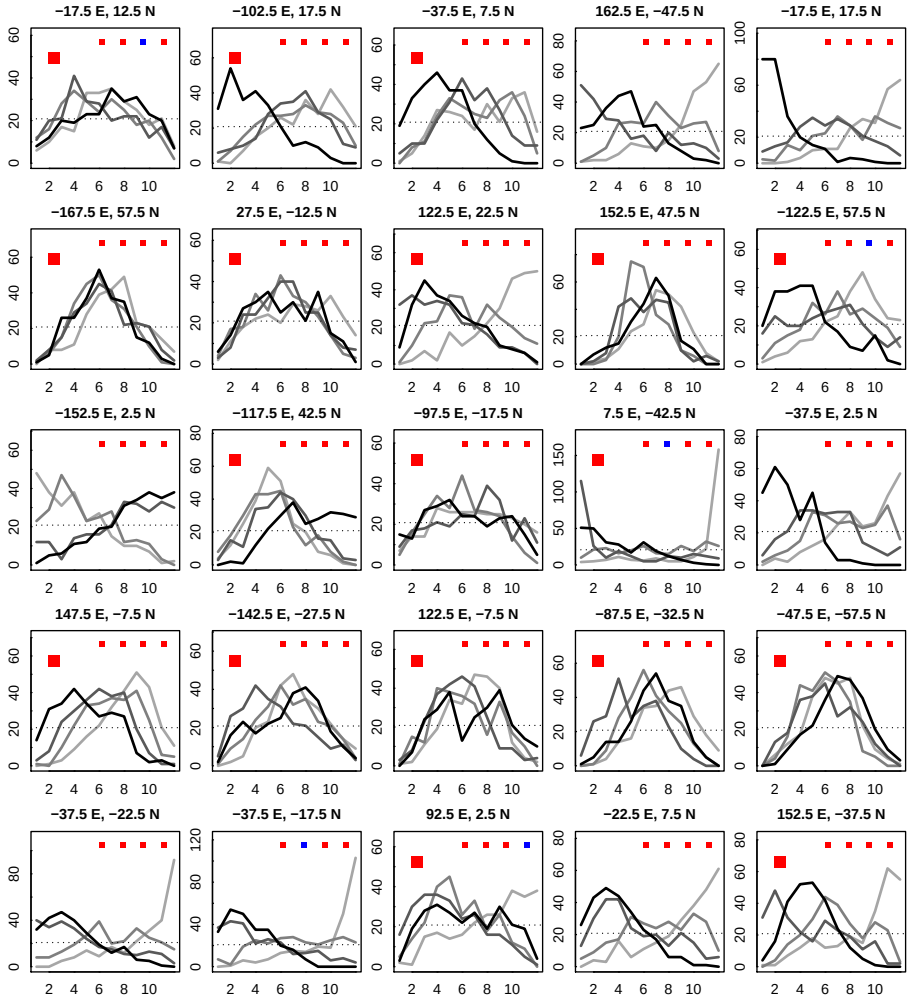


Fig. 10. Rank histogram counts for a random selection of 25 grid points from the decadal smooth global temperature data and the first, second, third and last 250 non-overlapping records of the decadally smoothed annual data (grey to black lines, about 800 to 1050, 1050 to 1300, 1300 to 1550, and 1595 to 1845). 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~~indicate non-significant deviations. Squares in each panel from left to right for the first, second, third and last sub-period. Locations given in titles of individual panels.

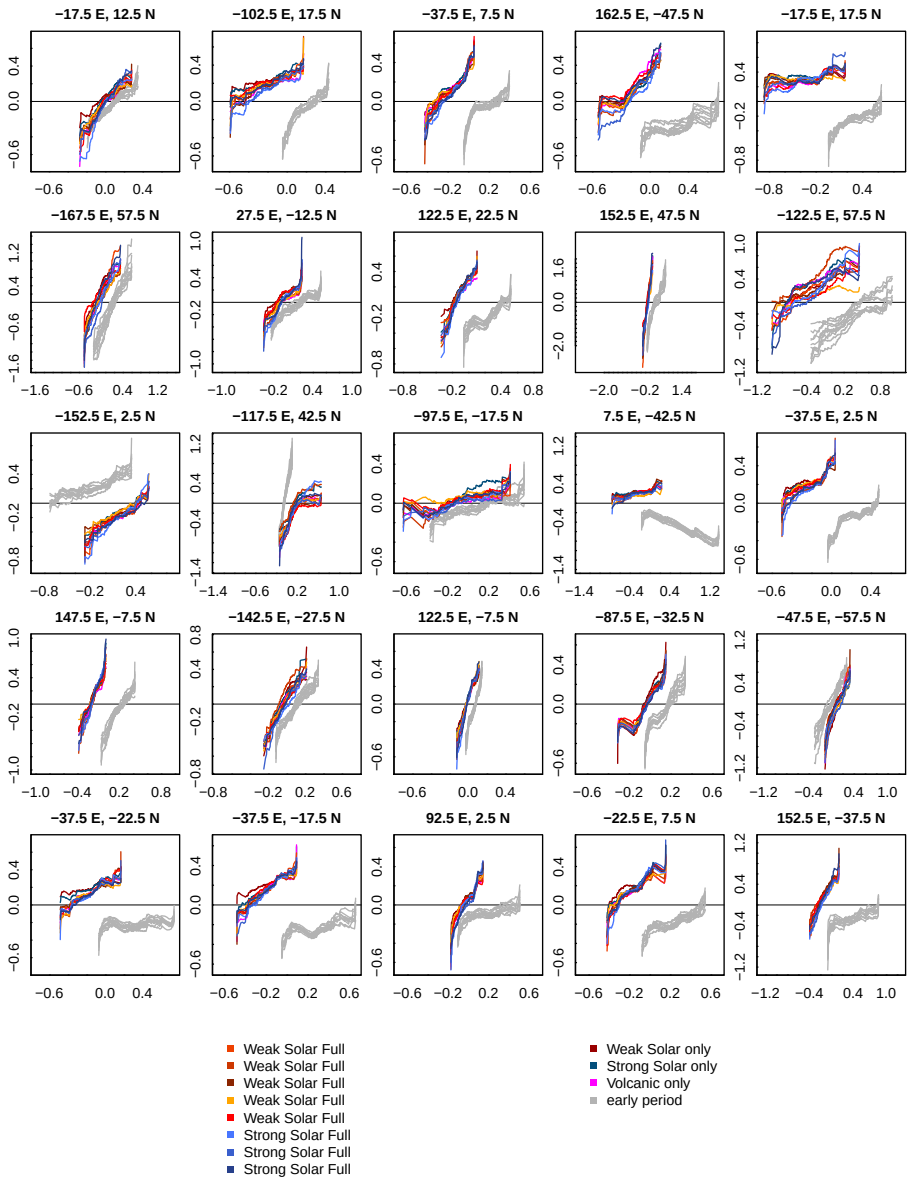


Fig. 11. Residual quantile-quantile plots for a random selection of 25 grid points from the decadal smooth global temperature data and the first (grey) and the last (~~colors~~colours) 250 records. Locations given in titles of individual panels. ~~Representation as in Fig. 3a~~For representation see legend.

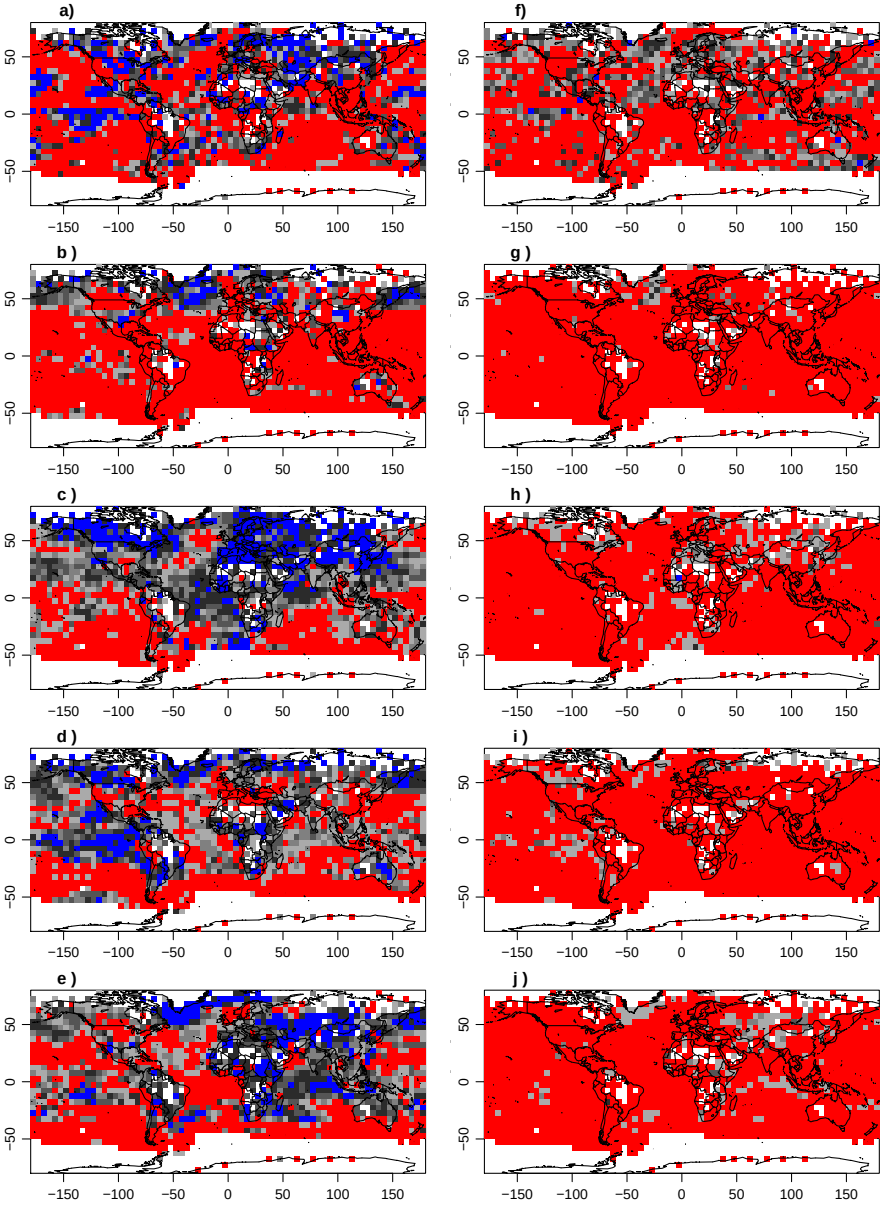


Fig. 12. Global assessment of the goodness-of-fit test for the decadal smooth data considering ~~errors~~ uncertainties in the ~~truth~~target. Plotted are lower p-values. In the left column: full χ^2 test, in the right column: maximum of p-values for single deviation tests for bias and spread. Blue smaller than 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.